Der Phish und die Nutzerin: Formalisierung, Konzeption und Analyse

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat)

am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Vorgelegt von
Oliver Wiese

Berlin 2023

Erstgutachter: Prof. Dr-Ing. Volker Roth
Zweitgutachter: Prof. Dr. Sven Dietrich

Tag der Disputation: 4. Juli 2023

Selbstständigkeitserklärung

Name: Wiese
Vorname: Oliver

Ich erkläre gegenüber der Freien Universität Berlin, dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die vorliegende Arbeit ist frei von Plagiaten. Alle Ausführungen, die wörtlich oder inhaltlich aus anderen Schriften entnommen sind, habe ich als solche kenntlich gemacht. Diese Dissertation wurde in gleicher oder ähnlicher Form noch in keinem früheren Promotionsverfahren eingereicht.

Mit einer Prüfung meiner Arbeit durch ein Plagiatsprüfungsprogramm erkläre ich mich einverstanden.

Datum: 9.Mai 2023 Unterschrift:

Zusammenfassung

Phishing-Angriffe sind im Internetzeitalter eine dauerhafte Erscheinung geworden. Eine nur technische Lösung zur Phishing-Bekämpfung bei E-Mails wurde bisher noch nicht gefunden. Stattdessen sind Nutzerinnen häufig gefordert die Angriffe selbst abzuwehren und werden hierfür geschult. In dieser Arbeit wird ausgehend von einer Beschreibung aktueller Phishing-Angriffe eine formale Modellierung zur Sicherheitsanalyse von Sicherheitsverfahren in der Darstellung von Phishing-Angriffen entwickelt. Diese Modellierung bildet eine Brücke zwischen den Formalismen aus der Sicherheitsforschung und der Kognitionswissenschaft.

Die Gefahren bei der Darstellung von E-Mail-Adressen werden durch Erkenntnisse der Kognitionswissenschaft empirisch gezeigt. Damit werden allgemeine Heuristiken zur Erkennung von möglichen Gefahrenquellen bei der Darstellung abgeleitet.

Ausgehend von der kognitiven Eigenschaft zur guten Wiedererkennung von Gesichtern wird ein Verfahren zur Darstellung von Absenderinnen vorgestellt und analysiert. Die Arbeit bietet damit eine neue Perspektive auf Phishing-Angriffe und menschen-zentrierte Sicherheit im Allgemeinen.

Danksagung

Nach langer und intensiver Arbeit ist meine Dissertation fertig! Diese Dissertation wäre ohne die ganz unterschiedliche und vielfältige Unterstützung von vielen verschiedenen Menschen nicht in dieser Form entstanden. An dieser Stelle möchte ich diese Unterstützung hervorheben und mich gleichzeitig bedanken.

An erster Stelle möchte ich mich bei meinem Doktorvater Volker Roth für die langjährige Unterstützung bedanken. Der Besuch der Rechnersicherheitsvorlesung im 4. Bachelorsemester hat mich motiviert und inspiriert, in diesem Forschungsfeld aktiv zu werden. Die gemeinsame Zusammenarbeit begann im Rahmen meiner Bachelorarbeit und führte schlussendlich zu dieser Dissertation. Durch die vielen Diskussionen, Anregungen und gemeinsamen Veröffentlichungen sowie die Finanzierung über diese lange Zeit wurde diese Arbeit erst ermöglicht, gefördert und unterstützt. Gleichzeitig wurden mir sowohl in der Lehre als auch in der Wissenschaft genügend Freiräume geboten, um den eigenen Forschungsinteressen zu folgen. Meinen Kolleginnen aus der Arbeitsgruppe gebührt ebenso mein Dank für die vielen Gespräche und gemeinsamen Arbeiten. Insbesondere bei Benjamin Güldenring, Joscha Lausch, Jakob Bode, Jan-Ole Malchow und Mani Esmaeili möchte ich mich für die gemeinsamen Arbeiten und zielführenden Diskussionen bedanken. Ohne die wertvolle organisatorische Unterstützung von Isabella Bargilly und Anna-Maria Hengst während meiner Zeit am Lehrstuhl hätte ich weniger Zeit für die Arbeit gehabt. Vielen Dank für eure Unterstützung!

Neben den Kolleginnen der Arbeitsgruppe Sichere Identität haben mich viele ehemalige Tutorinnen des Instituts für Informatik auch in den Jahren meiner Dissertation begleitet, unterstützt und motiviert. In alphabetischer Reihenfolge möchte ich mich darum bei Alexander Steen, Chris Pockrandt, Jonas Cleve, Kristin Knorr, Marcel Ehrhardt, Max Willert, Max Wisniewski, Nadja Seiferth bedanken. Daneben möchte ich mich noch bei Lydia Krause und Tobias Fiebig für den regen Austausch am Anfang meiner wissenschaftlichen Arbeit bedanken.

Daneben möchte ich mich bei Ingmar Camphausen und Bodo Riediger-Klaus für die technische Unterstützung im Rahmen meiner Dissertation bedanken.

Während meines Forschungsaufenthalts an der University of British Columbia haben mich die Gespräche und Arbeiten mit Kosta Beznosov, Artemij Voskobojnikov, Masoud Koushki und Borke Obada zusätzlich inspiriert und motiviert. Auch bei ihnen möchte ich mich bedanken.

Die gemeinsamen Forschungsprojekte mit Sascha Fahl, Christian Stransky und Yasemin Acar von der Universität Hannover haben meine Fähigkeiten im wissenschaftlichen Arbeiten zusätzlich gestärkt. Beispielsweise habe ich die Methoden zur Übertragung von Messergebnissen in ein Latex-Dokument adaptieren können. Auch ihnen möchte ich daher an dieser Stelle danken.

Ich möchte mich bei meiner Promotionskommission bestehend aus Larissa Zech, Sven Dietrich und Lutz Prechelt für die zügige und unkomplizierte Durchführung bedanken. Lutz Prechelt gebührt mein Dank auch für die Unterstützung zur Erlangung einer Zwischenfinanzierung während der Anfertigung dieser Dissertation.

Ein ganz besonderer Dank gilt Karin Wolff und Angelika Scharf für das Korrekturlesen der gesamten Arbeit. Ohne eure mühsame und ausführliche Arbeit wäre diese Dissertation nicht in diesem (grammatikalisch korrektem) Zustand! Daneben möchte ich mich bei Benjamin Güldenring, Alexander Steen, Manuel Riel und Jonas Winkler für das Lesen und Kommentieren von vorherigen Versionen der Dissertation bedanken.

Schlussendlich möchte ich mich bei meinen Eltern Martina Wiese und Rainer Wiese, meinem Bruder Sebastian Wiese, meinen Ruderkameradinnen und Mitgliedern der Steinbeißer-Klettergruppe für die vielfältige Unterstützung, Motivation und Begleitung über die gesamte Zeit der Dissertation bedanken. Die Gespräche und Unternehmungen außerhalb der Forschungsarbeit haben mir hierfür neue Kraft gegeben.

Ein besonderer Dank richtet sich an meine Frau Jacqueline Knoll, welche mich durch die schwierigen Phasen bei der Dissertation begleitet hat, mich dann immer motiviert hat und nach Möglichkeit unterstützt hat. Mit ihr konnte ich immer die schönen Erfahrungen der Dissertation feiern und in der vielen gemeinsamen Zeit neue Kraft tanken. Die (grammatikalischen) Korrekturen, vor allem in den letzten Zügen, waren auch immer sehr hilfreich.

Herzlichen Dank an alle Personen für eure Unterstützung!

Oliver Wiese

Inhaltsverzeichnis

1 Warum sich mit Phishing beschäftigen? 9

1.1 Beiträge . . . 11

1.2 Aufbau . . . 11

2 Hintergrund 13

2.1 Die E-Mail . . . 13

2.1.1 Versand und Empfang . . . 14

2.1.2 Sicherheitsprotokolle . . . 17

2.1.3 Zusammenfassung . . . 23

2.2 Nutzung von E-Mails . . . 24

2.2.1 Private Kommunikation . . . 26

2.2.2 Berufliche Kommunikation . . . 27

2.2.3 Zusammenfassung . . . 27

3 Die Kunst vom Angeln 28

3.1 Gefahrenmodell . . . 28

3.2 Fallbeispiele . . . 29

3.2.1 Massenangriffe . . . 29

3.2.2 Gezielte Angriffe . . . 30

3.2.3 Gemeinsamkeiten . . . 33

3.3 Abgrenzung zu anderen Angriffen . . . 35

3.4 Analysen zu Phishing in freier Wildbahn . . . 36

3.5 Zusammenfassung . . . 37

4 Formale Modellierung 39

4.1 Legitimitätsspiel . . . 39

4.1.1 Nutzerin $\mathcal{H}$ . . . 42

4.1.2 Erzeugung vom Kontext . . . 43

4.1.3 Sicherheitsspiel als Experiment . . . 46

4.2 Spiel mit Herkunft und Aktion . . . 49

4.3 Herkunft der Nachricht . . . 53
4.4 Einschränkung auf die Herkunftsdarstellung . . . 54
4.5 Einschränkungen . . . 57
4.5.1 Keine konkrete Umsetzung . . . 58
4.5.2 Schwächen außerhalb vom Modell . . . 58
4.5.3 Keine bedingungslose Sicherheit . . . 59
4.6 Zusammenfassung . . . 60
4.6.1 Verkettung von Sicherheitsspielen . . . 61

5 Gefährliche Zeichenketten 64
5.1 Identische Täuschungen . . . 64
5.1.1 Bekannte Fallstricke . . . 65
5.2 Fehlerhafte Wahrnehmung . . . 67
5.2.1 Buchstaben-Substitution . . . 68
5.2.2 Wahrnehmungskollision . . . 69
5.2.3 Weitere Fallstricke . . . 72
5.3 Unterschiedliche Interpretationen . . . 74
5.4 Feldbeobachtungen . . . 76
5.4.1 Methode . . . 76
5.4.2 Durchführung . . . 78
5.4.3 Ergebnis . . . 79
5.4.4 Diskussion . . . 82
5.4.5 Einschränkungen . . . 86
5.5 Verwandte Arbeiten aus der Sicherheitsforschung . . . 87
5.6 Zusammenfassung . . . 89

6 Sicherheit durch Wiedererkennung 91
6.1 Hintergrund und verwandte Arbeiten . . . 91
6.2 Verfahren und Darstellungsraum . . . 98
6.3 Einbettung im Sicherheitsspiel . . . 99
6.4 Abschätzung und Optimierung der Fehler . . . 103
6.4.1 Beschreibung vom Datensatz . . . 103
6.4.2 Modellauswahl . . . 104
6.4.3 Ergebnis . . . 105
6.4.4 Diskussion . . . 108
6.4.5 Sicherheitsabschätzung . . . 112
6.4.6 Einschränkungen . . . 114
6.5 Umsetzung . . . 115
6.6 Zusammenfassung . . . 116

7 Einordnung in den aktuellen Stand der Forschung **117**
7.1 Meta-Forschung 117
7.2 Warnungen 119
7.3 Verwandte formale Methoden 120
7.4 Ähnliche Verfahren 124

8 Schluss **125**
8.1 Ausblick 128
8.1.1 Verbesserung vom Formalismus 128
8.1.2 Formalisierung der menschen-zentrierten Sicherheit 129
8.1.3 Varianten vom Verfahren und andere Verfahren 129
8.1.4 Bedeutung vom Adressbuch 130
8.1.5 Sammlung von Felderfahrungen 130

Glossar **132**

Literaturverzeichnis **134**

A Über den Autor **146**

B Zusammenfassung der Ergebnisse **148**

C Reproduktion der Forschungsergebnisse **149**

Kapitel 1

Warum sich mit Phishing beschäftigen?

Beinahe jede Person mit einem E-Mail-Account kennt E-Mails mit der Aufforderung die Zugangsdaten für eine Bank auf einer Webseite einzugeben. Diese E-Mails geben vor, von der entsprechenden Bank zu sein. Dabei sind diese E-Mails oftmals nicht von der Bank, sondern von einer bösartigen Partei, welche die Zugangsdaten stehlen möchte. Diese und ähnliche Angriffe werden als Phishing-Angriffe bezeichnet. Bereits 2006 beantworteten Dhamija et al. die Frage, warum Phishing funktioniert [38]. Trotzdem ist Phishing immer noch präsent in unserem Alltag. Aus einer bürgerlich, westlichen Perspektive mögen die selbsterfahrenen Phishing-Angriffe einfältig, naiv und trivial wirken. Dabei schwingt immer die Frage mit: Sind solche Angriffe wirklich erfolgreich? Oder provokativer: Wer würde glauben, dass diese E-Mail von der Bank ist, und auf der Webseite die Zugangsdaten eingeben? Im Kapitel 4 wird hingegen deutlich, dass Phishing einerseits ein massives Problem ist, es vielfältige Zielgruppen gibt und diese erfolgreich angegriffen werden. Zu diesen Zielgruppen zählen unter anderem Personen, die sich in Autokratien für eine Zivilgesellschaft engagieren, und erfolgreiche Angriffe können schwere Konsequenzen nach sich ziehen. Häufig starten diese Angriffe mit einer E-Mail. Gleichzeitig starten Ransomware-Angriffe[1] ebenfalls mit einer E-Mail. Phishing und ähnliche Angriffe sind abseits von den persönlichen Erfahrungen von größerer Relevanz und eben nicht trivial. Aus einer akademischen Perspektive ist Phishing interessant, weil die Sicherheit aus einer sehr technischen Perspektive betrachtet werden kann, und gleichzeitig müssen die Darstellung sowie Interaktion der Menschen berücksichtigt werden. Aus der sehr technischen Perspektive können Phishing-Angriffe durch sichere (Internet)-Protokolle oder andere Maßnahmen erschwert werden. Beispielsweise konnte zu Anfangszeiten des Internets eine beliebige E-Mailadresse als `FROM`-Feld, also als Absenderin[2], angegeben werden und die E-Mail wurde zugestellt. In Kapitel 2 wird

[1] Als Ransomware werden Angriffe bezeichnet, bei denen Dokumente und andere Dateien von einer Angreiferin verschlüsselt werden und der Schlüssel zur Entschlüsselung gegen Zahlung eines Geldbetrages mitgeteilt wird [63].

[2] In dieser Arbeit wird die weibliche Form als generisches Femininum verwendet. Damit sind alle Menschen gemeint und inkludiert.

deutlich, dass dies praktisch nicht mehr möglich ist und eine gute IT-Hygiene von Servern dies verhindern kann. Doch damit haben Phishing-Angriffe nicht aufgehört und ein wesentlicher Aspekt dabei ist, dass bei einem Phishing-Angriff der Mensch die endgültige Entscheidung trifft. Aus diesem Grund muss bei einem Angriff eben nicht der Computer getäuscht werden, sondern Menschen mit ihren ganz eigenen Fähigkeiten sowie Stärken und Schwächen. Damit ist die Erkennung von Phishing-Angriffen und die Vermeidung von erfolgreichen Angriffen eine naheliegende und exzellente Forschungsaufgabe für die menschen-zentrierte Sicherheitsforschungsgemeinschaft. Diese Forschungsströmung begleitet dieses Thema nun bereits seit fast 20 Jahren. Die Forschungsschwerpunkte waren:

1. Wer ist eigentlich besonders anfällig für Phishing-Angriffe? [79, 120, 123]
2. Wie können Menschen zur Erkennung von Phishing-Angriffen geschult/trainiert/ausgebildet werden? [26, 77, 121, 143]
3. Wie können Menschen besonders effektiv vor Phishing-Angriffen gewarnt werden? [43, 79, 99]

Dagegen gibt es nur sehr wenige Vorschläge, wie eine grafische Oberfläche zur Darstellung einer Nachricht zu gestalten ist. Franz et al. haben durch eine Literaturrecherche mehr als 2000 Veröffentlichungen gefunden, aus denen sie 64 mit menschen-zentrierten Interventionen gegen Phishing identifiziert haben, und nur zwei dieser Veröffentlichungen haben Designvorschläge für die grafische Oberfläche von E-Mail-Anwendungen gemacht [47]. Verbesserungen in der grafischen Oberfläche können nicht nur Angriffe verhindern, sondern im Alltag die Benutzbarkeit erhöhen, wenn die Erkennung der Herkunft einfacher und schneller möglich ist. Die Konstruktion von einer sicheren Darstellung einer Herkunft einer Nachricht ist somit besonders erfolgversprechend, wurde aber in der Vergangenheit kaum verfolgt. Eine mögliche Ursache ist, dass es für die Evaluierung abseits von Studien keine geeigneten Methoden gibt. Die Studien in diesem Bereich sind komplex, aufwändig und immer anfechtbar. Eine Alternative aus der Informatik, insbesondere aus der theoretischen Informatik, ist die Entwicklung von Modellen und Formalismen als Abbild der Realität und der Möglichkeit einer (ersten) Evaluation auf Basis der entwickelten Abstraktion. Aus dieser Erkenntnis und Motivation leiten sich die folgenden Leitfragen für die nachfolgenden Kapitel ab:

1. Wie kann ein formales Modell zur Analyse von Phishing aussehen?
2. Wie kann eine Darstellung zur Erkennung von Angriffen konstruiert und evaluiert werden?

Das Ziel ist es, menschen-zentrierte-Sicherheit als eine Verknüpfung zwischen der Psychologie bzw. der Kognitionswissenschaften und der IT-Sicherheit bzw. der Kryptographie darzustellen und so der Forschungsgemeinschaft einen konstruktiven Ansatz anzubieten. Bei der Konstruktion eines Verfahrens ist das Ziel, Benutzbarkeit und Sicherheit nicht als gegensätzliches Spannungsfeld zu begreifen. Im Idealfall wird die Benutzbarkeit durch die Sicherheit auch erhöht und umgekehrt. Um diese Ziele zu erreichen, sind Einschränkungen nötig. Das Modell und die Konstruktionen von einer sicheren Darstellung ist eine theoretische Auseinandersetzung und lässt einen Gestaltungsspielraum für die Praxis. Es ist nicht der Anspruch, eine fertige Implementation von einer grafischen Oberfläche bereitzustellen, welche von Google,

Apple, Microsoft und anderen Konzernen einfach übernommen werden könnte und sofort alle Nutzerinnen vor Phishing-Angriffen schützt. Alle Unternehmen haben ihre eigenen Design-Richtlinien sowie Paradigmen. Eine einheitliche Lösung wird darum nicht möglich sein, sondern die Umsetzung ist immer in eine Anwendung eingebettet. Der Anspruch ist es, konstruktive Konzepte zu erarbeiten, welche von Entwicklerinnen dann in ihren Anwendungen aufgegriffen werden können.[3] Dies bedeutet, dass die Industrie aufgefordert ist, ihre Produkte wissenschaftlich zu evaluieren. Der Anspruch an dem Modell ist es, konstruktive Vorschläge aus der Wissenschaft zu evaluieren, die Industrie zu motivieren, sichere Vorschläge zu adoptieren und im Anschluss zu evaluieren. Im Idealfall werden bei der Analyse im Modell mögliche Fallstricke und Gefahren, welche bei der praktischen Umsetzung zu berücksichtigen sind, hervorgehoben.

1.1 Beiträge

Die Beiträge dieser Arbeit sind zusammengefasst wie folgt:

1. Es wird ein formales Modell entwickelt, welches die Sicherheit gegen Phishing-Angriffe einer Darstellung einer E-Mail in Beziehung zu kognitiven Eigenschaften von Menschen stellt. In diesem Modell werden die Darstellung und die menschliche Entscheidung abstrakt betrachtet. Dies ermöglicht eine Herleitung und Begründung der Sicherheit gegen Phishing-Angriffe.
2. Es werden Heuristiken zur frühzeitigen Erkennung von potentiell unsicheren Darstellungen von E-Mails aufbauend auf kognitiven Einschränkungen von Menschen entwickelt. In einer Datenanalyse von tatsächlichen Phishing-E-Mails werden diese Heuristiken auf E-Mail-Adressen angewendet und die Beziehung zu Einschränkungen von Menschen beim Lesen von Zeichenketten zeigt.
3. Es wird eine Darstellung einer Herkunft einer E-Mail mittels zufällig gewählter Gesichter vorgeschlagen und die Sicherheit dieser Darstellung durch die Ergebnisse aus Experimenten der Kognitionswissenschaft begründet.

Die letzten beiden Beiträge zeigen dabei die Anwendbarkeit vom ersten Beitrag und ergänzen sich damit.

1.2 Aufbau

Die nachfolgenden Kapitel sind wie folgt aufgebaut. Im nächsten Kapitel wird der Hintergrund zur E-Mail und deren Nutzung aufgezeigt. Im Anschluss folgt ein Kapitel zum Gefahrenmodell von Phishing und Beobachtungen von Phishing in der Praxis. Im vierten Kapitel wird ein formales Modell von Phishing entwickelt und vorgestellt. Dieses formale Modell und kognitive Eigenschaften von Menschen beim

[3] Grundsätzlich sollte sich die akademische Forschungsgemeinschaft die Fragen stellen, ob sie (Industrie)-Design für Unternehmen anbieten sollten und es überhaupt kann. Denn sowohl bei Industrieprodukten als auch bei Webseiten, Apps und anderen IT-Anwendungen hat sich eine Arbeitsteilung zwischen Entwicklerinnen/ Ingenieurinnen und (Industrie/Web/App)-Designerinnen gebildet. Diese ist mit den (beschränkten) Ressourcen in der akademischen Wissenschaft selten gegeben.

Lesen von Zeichenketten werden im fünften Kapitel benutzt, um die Unsicherheit durch E-Mail-Adressen hervorzuheben. Mit dieser Motivation wird im darauffolgenden Kapitel ein Verfahren aufbauend auf Erkenntnissen der Kognitionswissenschaft konstruiert und analysiert. Im Anschluss werden die drei Beiträge mit der bestehenden Literatur verglichen. Der Schluss fasst die Ergebnisse kurz zusammen und gibt einen Ausblick auf weitere zukünftige Forschungsmöglichkeiten.

Kapitel 2

Hintergrund

In diesem Kapitel werden die E-Mail, deren Nutzung und Gefahren erläutert und diskutiert. Es wird die technische Komplexität, als auch der Gestaltungsraum für Maßnahmen gegen Phishing dargestellt. Es werden zunächst der Aufbau einer E-Mail mit den wichtigsten Protokolle vorgestellt. Im Anschluss wird die Nutzung der E-Mail im privaten, wie beruflichen Kontext erläutert.

2.1 Die E-Mail

Eine E-Mail ist eine menschenlesbare Datei, welche sich in Meta-Daten (dem `Header`) und Inhaltsdaten unterteilt, die zwischen den Geräten, Anwendungen und Servern ausgetauscht werden. Abbildung 2.1 zeigt eine einfache E-Mail im Texteditor und in einer Endanwendung. Eine E-Mail im Texteditor ist für Menschen lesbar und nachvollziehbar, aber sehr technisch. Es beinhaltet deutlich mehr Informationen als in der Endanwendung. Beispielsweise ist die verwendete Anwendung beim Versand erkennbar. Die E-Mail kann aber deutlich komplexer werden und noch mehr Informationen beinhalten.

Diese werden durch verschiedene Spezifikationen (RFC) definiert und beschrieben. Die Anzahl der Spezifikationen ist historisch gewachsen und Spezifikationen wurden vereinzelt weiterentwickelt. Einer der

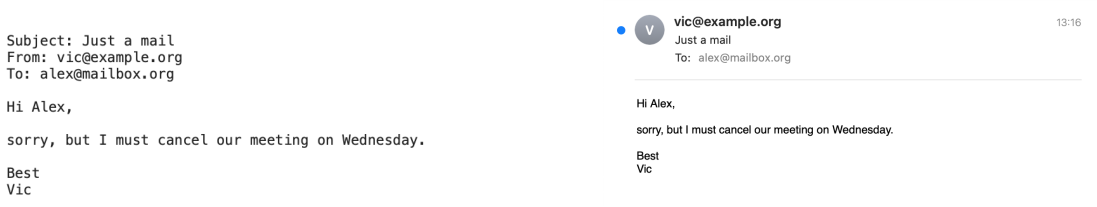

Abbildung 2.1: Eine simple E-Mail zwischen zwei Personen. Links wird die E-Mail im Texteditor dargestellt und rechts in der AppleMail Anwendung auf MacOSX.

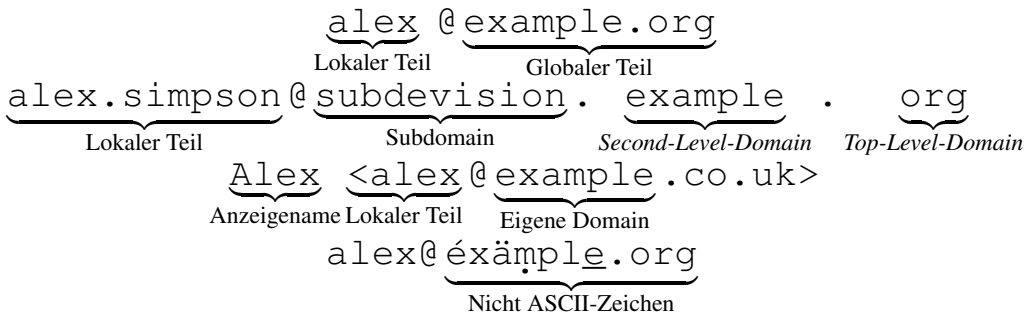

Abbildung 2.2: Schematische Darstellungen von E-Mail-Adressen und deren Bedeutungen. Eine *Second-Level-Domain* muss nicht zwangsläufig der Name der eigenen Domain sein. Zusätzlich kann noch ein Anzeigename angegeben werden und Zeichen können Nicht-ASCII-Zeichen sein.

ersten RFCs zur der E-Mail wurde 1973 veröffentlicht [15]. Dennoch ist der allgemeine Ablauf beim Versand und Empfang einer E-Mail ähnlich geblieben.

Ein wichtiger Aspekt im Kontext der E-Mail ist die Adresse. Sie gibt Auskunft über Absenderin und Empfängerin einer E-Mail. Die Adresse unterteilt sich in einen globalen Teil, welche den öffentlich erreichbaren Server benennt, und den lokalen Teil, welcher es dem eingehenden Server und Dienstleister ermöglicht, die E-Mail der richtigen Mailbox und damit der richtigen Person zuzustellen. Beide Teile werden durch das `@`-Zeichen getrennt und obliegen genauen Spezifikationen, welche Zeichen in den entsprechenden Teilen zugelassen sind. Der globale Teil umfasst mindestens eine *Top-Level-Domain* und eine *Second-Level-Domain*. Er kann um weitere *Subdomains* ergänzt werden. In manchen Fällen ist die Registrierung der eignen Domain nur als *Subdomain* möglich. Dies betrifft zum Beispiel Domains mit der Endung `co.uk`. Daneben kann die E-Mail-Adresse einen Anzeigenamen umfassen. Dieser kann frei gewählt werden. Abbildung 2.2 stellt verschiedene mögliche E-Mail-Adressen dar.

Die E-Mail wurde im RFC5322 [111] spezifiziert, aber auch durch weitere RFCs ergänzt. Anfangs waren nur ASCII-Zeichen zugelassen, aber durch RFC6533 [86] wurden auch andere Zeichensätze und Alphabete im Sinne einer Internationalisierung zu gelassen.

2.1.1 Versand und Empfang

Beim Austausch einer Nachricht gibt es verschiedene Parteien. Die wichtigsten sind die Absenderin, die Empfängerin sowie die Server des jeweiligen E-Mail-Providers. Ein E-Mail-Provider ist die Anbieterin von Servern, welche es einer Person erlaubt, E-Mails zu versenden und zu empfangen. Im privaten Kontext kann eine Person einen kommerziellen E-Mail-Provider frei wählen oder selber die Server hosten. Im beruflichen Kontext wird der E-Mail-Provider von der Organisation bestimmt. Entweder kann die Organisation die Infrastruktur selber bereitstellen oder auf einen kommerziellen Anbieter zurückgreifen. Ein vereinfachter Ablauf ist schematisch in Abbildung 2.3 dargestellt.

Die Absenderin Vic verfasst ihre Nachricht mit Hilfe einer E-Mail-Anwendung. Typischerweise ist dies

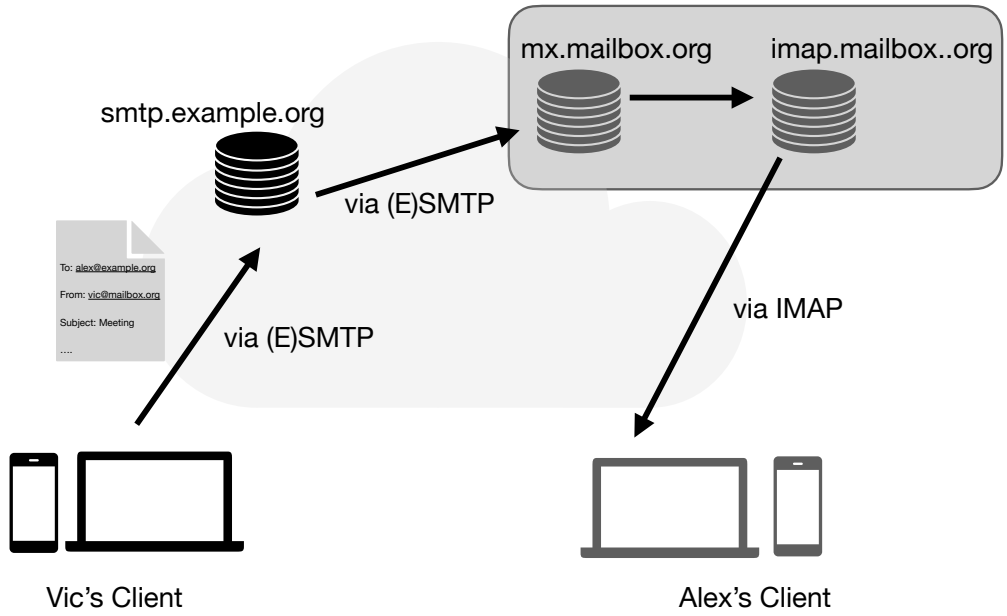

Abbildung 2.3: Versand und Empfang einer E-Mail. Vic möchte eine Nachricht über das nächste Meeting an Alex versenden.

entweder eine Anwendung auf einem Laptop, Desktop-Rechner, Smartphone oder eine Webseite im Browser. Formal wird diese Anwendung als *Mail User Agent* (MUA) bezeichnet. Die eigentliche E-Mail ist eine Textdatei, welche während der Zustellung erweitert wird. Die E-Mail wird mittels dem Protokoll SMTP (*Simple Mail Transport Protocol*) [70] an den SMTP-Server vom E-Mail-Provider der Absenderin übertragen. Der *Mail Transport Agent* (MTA) sucht anschließend nach dem DNS-Eintrag vom empfangenden Server. In dem DNS-Eintrag ist der Empfangsserver für eine E-Mail eingetragen und an diesen wird die E-Mail mittels SMTP zugestellt. In unserem Beispiel ist das mx.mailbox.org. Diese leitet die E-Mail meist in einem internen Netz an einen Server weiter. Die Empfängerin kann die Nachricht mittels ihres Clients und dem Protokoll IMAP (*Internet Message Access Protocol*) [87] von diesem Server abrufen.

In der Abbildung 2.3 werden nur wenige Server der Provider dargestellt. In der Praxis ist es durchaus üblich, dass eine E-Mail über verschiedene Server weitergereicht wird. Dies kann bei der E-Mail meist einfach nachvollzogen werden, denn die Server protokollieren den Empfang in der E-Mail und diese kann von Nutzenden gelesen werden.

Abbildung 2.4 zeigt einen Auszug aus den Meta-Daten einer E-Mail vom VDIVDE-IT an eine Adresse an der Freien Universität Berlin. In diesen Meta-Daten wird erkennbar, welche Server diese E-Mail erhalten haben und weitergeleitet haben. Die Server sind angehalten, den Erhalt einer E-Mail mit einem `Received`-Eintrag zu protokollieren. Auf der Seite des VDIVDE-IT sind zwei Server erkennbar

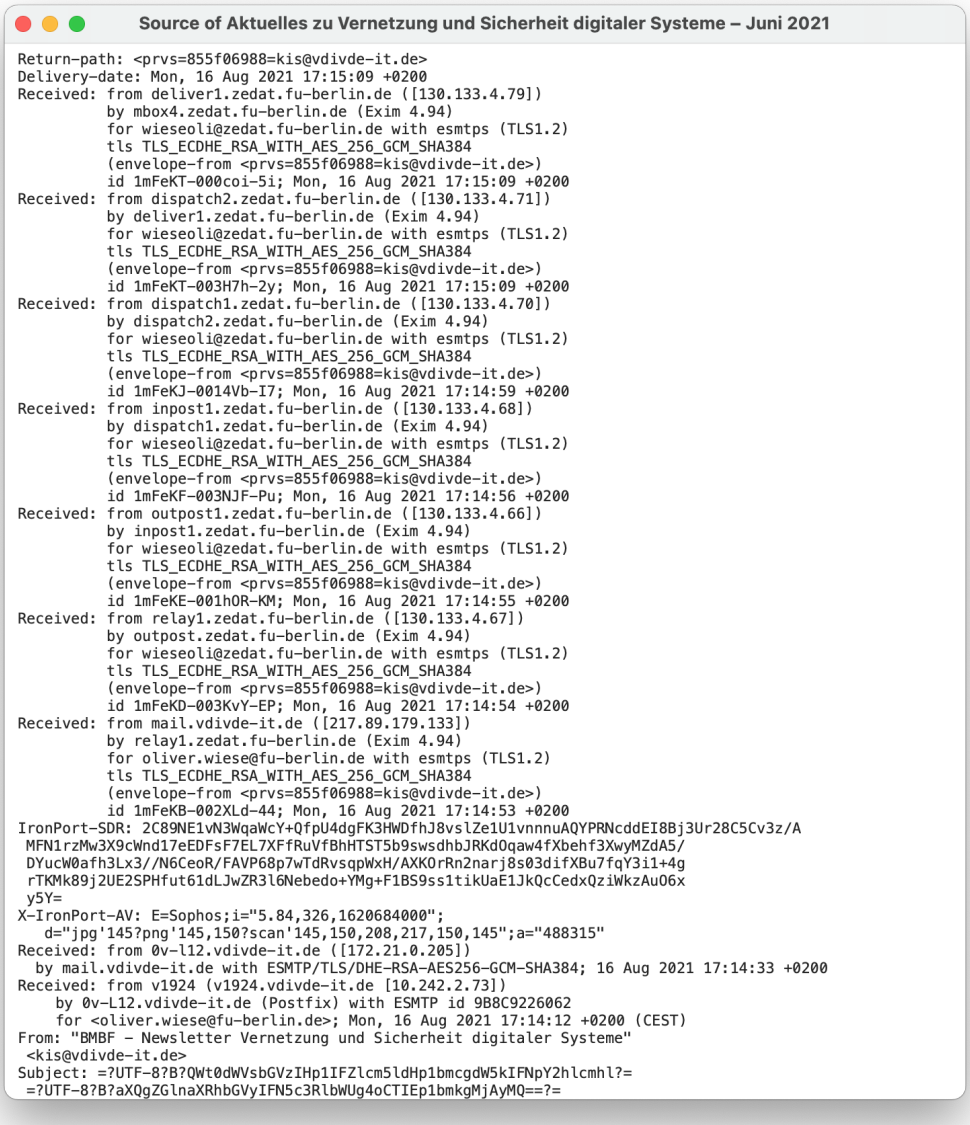

Abbildung 2.4: Die Meta-Informationen einer E-Mail sind für Nutzende verfügbar und einsehbar. Dieses Beispiel zeigt eine komplexere Abfolge von verschiedenen Servern an.

(`v1924.vdivde-it.de` und `0v-L12.vdivde-it.de`). Auf der empfangenden Seite der Freien Universität Berlin sind insgesamt sieben Server involviert. Dies zeigt die komplexen Strukturen beim Empfang und Versand einer E-Mail. Diese Eintragungen sind zunächst erstmal nicht kryptographisch geschützt. Sie können einfach editiert und manipuliert werden. Erst durch zusätzliche Protokolle kann die Integrität der `Received`-Einträge gewährleistet werden. Dies ist ein Phänomen, welches viele Sicherheitsaspekte einer E-Mail betrifft. Aus diesem Grund werden im Folgenden ausgewählte Zusatzprotokolle vorgestellt, welche die Sicherheit einer E-Mail erhöhen.

2.1.2 Sicherheitsprotokolle

Die Sicherheit einer E-Mail kann durch zusätzliche Protokolle in Bezug auf unterschiedliche Sicherheitseigenschaften erhöht werden. Die allgemeinen Sicherheitsziele einer E-Mail sind *Vertraulichkeit*, *Integrität*, *Authentizität* und unter Umständen *Nichtabstreitbarkeit* [80].

Definition 1. *[Vertraulichkeit] [80] Eine E-Mail ist vertraulich, wenn der Inhalt der E-Mail gegenüber Dritten geschützt ist.*

Im Kontext von E-Mail kann die Vertraulichkeit unterschiedlich ausgelegt werden. Es wird zwischen einer Vertraulichkeit während des Transportes einer Nachricht über das öffentliche Internet und der Vertraulichkeit gegenüber den Servern der Mail-Provider unterschieden. Letzteres umfasst insbesondere die Vertraulichkeit einer E-Mail, wenn diese auf dem Server gespeichert wird.

Die Vertraulichkeit einer E-Mail wird durch die Verschlüsselung des Inhalts gewährleistet [80]. Vertraulichkeit während des Transports zwischen verschiedenen Servern wird mittels Transportverschlüsselung erreicht. Die Vertraulichkeit gegenüber dem Server wird durch Ende-zu-Ende-Verschlüsselung erreicht und umfasst die Verschlüsselung und Entschlüsselung einer E-Mail auf den Endgeräten der sendenden und empfangenden Person, wobei der Schlüssel zur Entschlüsselung allen anderen unbekannt ist. Daneben kann die Vertraulichkeit auch durch Zugriffskontrolle geschützt werden werden.

Definition 2. *[Integrität] [80] Integrität bedeutet im Kontext der E-Mail, dass der Inhalt einer E-Mail so empfangen wurde, wie er versendet wurde.*

Die Integrität einer E-Mail kann durch *Message Authentication Codes* oder digitale Signaturen erreicht werden [80].

Definition 3. *[Authentizität] [80] Eine E-Mail ist authentisch, wenn die Herkunft einer E-Mail gegenüber der Empfängerin nachgewiesen werden kann.*

Die Herkunft einer E-Mail kann einerseits nur auf den E-Mail-Provider als sendendender Server reduziert werden. Andererseits kann die Herkunft auch auf die Person bezogen werden. Authentizität kann durch *Message Authentication Codes* oder digitale Signaturen erreicht werden, wenn die empfangende Seite überzeugt ist, dass die absendende Seite im Besitz eines bestimmten geheimen Schlüssels ist [80]. Dies bedeutet aber nicht, dass eine dritte Partei davon überzeugt werden kann. Dies wird mittels Nichtabstreitbarkeit definiert.

Definition 4. *[Nichtabstreitbarkeit] Eine E-Mail ist nicht abstreitbar, wenn die empfangende Seite eine dritte Partei von der Herkunft einer E-Mail überzeugen kann.*

Dies kann durch digitale Signaturen realisiert werden, indem die empfangende Seite eine dritte Partei überzeugen kann, dass nur die Besitzerin eines geheimen Schlüssels die E-Mail versendet haben kann [80].

Eine nicht abstreitbare E-Mail ist auch authentisch und die Integrität ist gewährleistet, aber die umgekehrte Richtung ist nicht immer garantiert [80]. Wie bereits bei der Vertraulichkeit einer E-Mail muss bei der Integrität, Authentizität und Nichtabstreitbarkeit einer E-Mail unterschieden werden, ob die Sicherheitseigenschaft bedeuten soll, dass die E-Mail über einen bestimmten E-Mail Provider versendet wurde oder von einer bestimmten Person. Im Folgenden werden verschiedene Protokolle zur Umsetzung von Sicherheitseigenschaften vorgestellt. Die Nutzung der Protokolle wurde in wissenschaftlichen Studien gemessen und untersucht [46, 42, 62, 128].

Transport-Verschlüsselung

Die Kommunikation zwischen Server und Client erfolgt mittels SMTP oder IMAP und kann mittels TLS oder STARTTLS geschützt werden [46, 42, 60, 96]. Gegenüber passiven Angriffen sind TLS und STARTTLS bei Nutzung von sicheren Algorithmen ausreichend [46]. Bei aktiven Angriffen kann die Angreiferin versuchen, sich als eine der beiden Parteien auszugeben und so die E-Mail über sich an die richtige Partei weiterleiten [46]. Damit die Angreiferin keinen Server nachahmen kann, muss die andere Partei das entsprechende Zertifikat verifizieren. Dies sind optionale Möglichkeiten, welche nicht von jedem Server umgesetzt werden müssen. Foster et al. haben 2014 und 2015 untersucht, ob E-Mail-Provider diese Sicherheitsfunktionen unterstützen [46]. Sie haben dazu 22 populäre E-Mail-Provider ausgewählt und deren Konfiguration getestet. Nur ein einziger Server unterstützte bei SMTP und IMAP kein TLS. Die Mehrheit bei SMTP (12 von 22) und bei IMAP (15 von 22) verweigerten Verbindungen ohne TLS. Allerdings war die Verifikation der Zertifikate nicht immer möglich, weil der Hostname vom Server nicht im Zertifikat stand. Dies betraf bei SMTP 6 von 22 Servern und bei IMAP 4 von 22 Servern. Foster et al. untersuchten, ob die E-Mail-Provider untereinander beim Versand die Verbindung mittels TLS schützen. [46] Dazu betrachten sie die `Received`-Einträge, wie beispielsweise in Abbildung 2.4 darstellt. Dort ist auch nachvollziehbar, ob eine Verbindung mit TLS geschützt war. In Abbildung 2.4 ist erkennbar, dass alle Übermittlungen mit TLS 1.2 geschützt waren. Sie konnten beobachten, dass häufig die Kommunikation zwischen den Server nicht geschützt war, aber es deutliche Verbesserungen zwischen 2014 und 2015 zu beobachten gab [46].

Durumeric et al. untersuchten die SMTP Log-Einträge von Gmail zwischen 2014 und 2015. [42] Sie konnten Unterschiede zwischen eingehenden und ausgehenden E-Mails erkennen. Am Ende der Analysezeit im April 2015 waren 80% aller ausgehenden E-Mails von Gmail mit STARTTLS geschützt, aber nur 60% der eingehenden E-Mails. Dennoch konnten sie Verbesserungen zum Beginn der Messung im Januar 2014 beobachten. Bei ausgehenden E-Mails waren anfangs nur 52% geschützt und bei eingehenden

nur 33% [42].

Gmail veröffentlicht eine Übersicht über mit TLS transportierten Nachrichten in ihrem Transparenzbericht [1]. Im Jahr 2022 wurden mehr als 92% aller ausgehenden E-Mails verschlüsselt und mehr als 95% aller eingehenden E-Mails mit TLS verschlüsselt.

Durumeric et al. konnten in der Datenanalyse der TLS Verbindungen mit Gmail keine Verwendung von verwundbaren Verschlüsselungsalgorithmen finden, wobei die Verwendung von schwachen Algorithmen beobachtet werden konnte [42].

Zusammenfassend ist zu beobachten, dass Transportverschlüsselung mittels TLS bei gängigen E-Mail-Providern verwendet wird und vor passiven Angriffen geschützt ist, wenn dem E-Mail-Provider vertraut wird und dieser nicht als spionierender Dritter betrachtet wird.

SPF

Der RFC7208 [69] spezifiziert das *Sender Policy Framework* (SPF). In diesem Protokoll kann ein E-Mail-Server die IP-Adressen, welche zum Versand von E-Mails für eine bestimmte Domain genutzt werden, über einen DNS-Eintrag veröffentlichen. Ein Server, der eine E-Mail von der entsprechenden Domain empfängt, kann dann überprüfen, ob die zum Versand verwendete IP-Adresse als Versandadresse publiziert wurde und somit E-Mails im Namen der Domain versenden darf [69].

Es erhöht damit die Authentizität einer E-Mail. Es schützt damit nicht die komplette E-Mail-Adresse, sondern nur den Domain-Teil, dient primär als Protokoll zwischen zwei Servern und ist kein kryptographisches Protokoll.

In der Studie von Foster et al. hatten 15 von 22 E-Mail-Providern und 42,26% von den Alexa top million Domains einen SPF Eintrag in ihrem DNS-Eintrag [46]. Von den 22 E-Mail-Providern lehnten nur fünf eine E-Mail von einem nicht autorisierten Server ab und fünf verschoben die E-Mail in den Spam-Ordner. Durumeric et al. berichten, dass Gmail bei 92% der eingehenden E-Mails erfolgreich SPF validieren konnte [42].

SPF erlaubt die Veröffentlichung von Sicherheitsrichtlinen zum Umgang mit E-Mails, welche von nicht autorisierten Servern versendet wurden. Der ausgehende E-Mail-Server kann neben den IP-Adressen diese Sicherheitsrichtline im DNS-Eintrag hinterlegen. Ein Server, welcher eine E-Mail von der entsprechenden Domain erhält aber von einer anderen IP-Adresse versendet wurde, hat somit eine Empfehlung zum Umgang mit diesen fragwürdigen E-Mails. Beispielsweise kann empfohlen werden, nicht autorisierte E-Mails zu löschen und nicht zuzustellen. Durumeric et al. berichten, dass 58% der Server empfehlen, die Nachricht als verdächtig zu markieren und zum Beispiel in den Spam-Ordner abzulegen. 21,7% der Server empfahlen, solche E-Mails nicht anzunehmen und die restlichen Server hatten keine Empfehlung [42].

[1] Siehe https://transparencyreport.google.com/safer-email/, letzter Zugriff: 25. April 2023

Im Gegensatz zu eingehenden Servern kann eine Endanwendung einen SPF Eintrag nur validieren, wenn die eingehende IP-Adresse vom fremden Server im `Received`-Feld hinterlegt ist. Die `Received`-Felder sind nicht kryptographisch gesichert und können damit manipuliert werden. Die Auswertung in einer Endanwendung ist damit eingeschränkt.

DKIM

Eine deutliche Verbesserung der Authentizität der Absenderin ist möglich, wenn der absendende Server eine E-Mail signiert und der empfangende Server diese Signatur prüfen kann. Dieses bedeutet zwar nicht, dass die Absenderin überprüfbar ist, aber der absendende Server ist authentisch überprüfbar.

Der RFC6376 [75] spezifiziert solch eine Möglichkeit und benennt dies als *Domain Key Identified Mail* (DKIM). Der absendende Server berechnet eine digitale Signatur über selbstgewählte Meta-Daten und dem Inhalt der E-Mail. Diese digitale Signatur wird mit weiteren Informationen, wie zum Beispiel verwendeten Algorithmus, den ausgewählten Feldern von der E-Mail als Meta-Daten hinzugefügt. Der passende öffentliche Schlüssel für die digitale Signatur wird anschließend im DNS-Eintrag hinterlegt und der empfangende Server kann diese überprüfen. Im Gegensatz zu SPF kann kryptographisch der absendende Server geprüft werden und bietet so mehr Sicherheit. Allerdings gilt die Authentizität der Absenderin zunächst nur für den Domain-Namen und nicht für die komplette Adresse. Ein E-Mail Server mit einer sicheren Konfiguration überprüft vor dem Versand, ob Nutzerinnen die angegebene E-Mail Adresse im From-Feld benutzen dürfen. Kaiwen et al. diskutieren dieses Problem und die Problematik bei der Weiterleitung einer E-Mail [119]. Denn nach dem DKIM Protokoll wird bei einer Weiterleitung eine DKIM Signatur hinzugefügt werden. Dies kann dafür sorgen, dass eine E-Mail ohne vorherige DKIM Signatur durch die Weiterleitung authentischer wird.

In der Studie von Foster et al. von 2014 hatten nur 13 von 22 E-Mail-Provider DKIM unterstützt [46]. Eine Regelung zur Behandlung von fehlerhaften Signaturen kann der absendende Server aber nicht festlegen. 83% der E-Mails an Gmail im April 2015 wurden mit DKIM signiert, wobei es eine Fehlerquote von 6,14% gab [42]. Diese hohe Fehlerquote ist ein möglicher Ansatz für die Beobachtung von Foster et al., dass in ihrer Studie nur 3 E-Mail-Providern bei einer fehlerhaften Signatur eine E-Mail gesondert behandelten oder ablehnten [46]. Der E-Mail-Provider lehnt eventuell eine wichtige E-Mail für Nutzerinnen ab und sorgt so für Frustration und Enttäuschung bei Nutzerinnen.

DMARC

Die Protokolle DKIM und SPF bestehen nebeneinander und ergänzen sich. Die Entscheidung über das Verhalten bei fehlerhaften E-Mails ist für einen empfangenden Server nicht einfach und ist stark abhängig von der Absenderin. Beispielsweise wird eine Online-Bank wahrscheinlich keine E-Mails fehlerhaft signieren. Bei einem Testsystem kann dies aber passieren. Damit der empfangende Server eine Entscheidungshilfe bekommt, wurden SPF und DKIM im RFC7489 [76] für *Domain-based Message Authentica-*

tion, Reporting, and Conformance (DMARC) gebündelt und die Absenderin kann eine Handlungsempfehlung im DNS-Eintrag hinterlegen.

Die Verbreitung von DMARC ist aber nur eingeschränkt. Nur 26,1% der eingehenden E-Mails von Gmail im April 2015 kamen von Servern mit einem DMARC-Eintrag und 1.1% der Alexa Top Millionen Domains mit einem MX-Eintrag hatten einen DMARC-Eintrag veröffentlicht [42]. Die meisten von diesen Einträgen waren aber leer und hatten damit keine Empfehlung für die empfangende Seite. In der Studie von Foster et al. empfahlen nur 3 von 22 Servern, ungültige Nachrichten zu verwerfen, und 10 Server hatten überhaupt nur einen DMARC-Eintrag [46].

S/MIME

Die bisherigen Sicherheitsprotokolle erhöhen die Sicherheit zwischen Servern und erfordern, dass dem Server vertraut wird. Sie stellen keine direkte Sicherheit zwischen der Absenderin und der Empfängerin her. Ende-zu-Ende-Sicherheit ermöglichen die zwei verschiedenen Sicherheitsprotokolle S/MIME und OpenPGP. Beide ermöglichen eine Verschlüsselung zwischen beiden Endgeräten und ermöglichen damit eine Ende-zu-Ende-Verschlüsselung. Zusätzlich bieten sie die Möglichkeit, E-Mails digital zu signieren.

Im RFC3211 [54] wird das Protokoll *Secure/Multipurpose Internet Mail Extensions* (S/MIME) spezifiziert und basiert auf einer (kommerziellen) Infrastruktur mittels öffentlicher Schlüssel (PKI). Hierzu erstellt sich eine Nutzerin ein Schlüsselpaar aus einem geheimen sowie einem öffentlichen Schlüssel und speichert dieses Schlüsselpaar lokal. Der öffentliche Schlüssel wird durch ein Zertifikat einer Zertifikatsautoritätsstelle (CA) autorisiert und legitimiert. Beispielsweise ist das deutsche Forscher Netz (DFN) Teil einer PKI und die Universitäten können ihren Studierenden und ihrem Personal Zertifikate ausstellen.

Wenn beide Parteien S/MIME nutzen, kann die Kommunikation Ende-zu-Ende verschlüsselt werden und bietet damit den höchsten Schutz der Vertraulichkeit. Hierzu muss den E-Mail-Servern nicht vertraut werden. Einzig der PKI muss vertraut werden, indem das richtige Zertifikat verwendet wird. Wenn die Absenderin S/MIME nutzt, kann sie ihre E-Mails signieren. Diese sind dann nicht-abstreitbar, denn nur die Besitzerin des geheimen Schlüssels konnte die E-Mail mit dem Zertifikat signieren. Dies ist die höchste Stufe der Authentizität einer E-Mail, denn es beweist nicht nur, dass ein bestimmter E-Mail-Provider die E-Mail versendet hat, sondern eine bestimmte Person oder Organisation bzw. Schlüsselbesitzerin. Im Gegensatz zu den vorherigen Protokollen müssen Nutzerinnen aktiv werden, eine entsprechende E-Mail-Anwendung benutzen und ein Zertifikat erstellen. In einer Analyse des gespeicherten E-Mail-Verkehrs einer deutschen Universität haben Stransky et al. [128][2] die Verbreitung von Ende-zu-Ende-Verschlüsselung untersucht. Dazu wurden die Meta-Daten von 81 Millionen E-Mails untersucht und unter 3% der E-Mails waren verschlüsselt oder signiert, wobei nur insgesamt 0,06% aller E-Mails verschlüsselt waren. Dies zeigt, dass die Nutzung von digitalen Signaturen und insbesondere von Ende-zu-Ende-Verschlüsselung nur ein Randphänomen ist. Nur 5,5% der Nutzenden der Universität verwendeten entweder OpenPGP oder

[2]Diese Studie war ein Kooperationsprojekt zwischen der Leibniz Universität Hannover und der Freien Universität Berlin unter meiner Beteiligung.

S/MIME, wobei nur 208 Personen mit S/MIME E-Mails verschlüsselten [128]. Demgegenüber ist die Nutzung von digitalen Signaturen mehr verbreitet. 601 Personen nutzten S/MIME zum Signieren einer E-Mail. Diese Personen signierten im Durchschnitt ein Drittel ihrer E-Mails, wohingegen im Durchschnitt unter 2% der E-Mails verschlüsselt wurden [128]. Im Gegensatz zu der Signierung von E-Mails erfordert die Verschlüsselung in der Regel, dass beide Personen S/MIME nutzen. Dennoch wurden selbst nur 3,36% der E-Mails zwischen zwei Personen mit einem S/MIME Zertifikat verschlüsselt [128]. Dies zeigt, dass Ende-zu-Ende Verschlüsselung bei der E-Mail kaum Anwendung findet; selbst dann, wenn es möglich wäre. Bei der Signierung hingegen wird deutlich, dass wenn nur eine E-Mail-Anwendung verwendet wird, fast zwei Drittel der E-Mails im Durchschnitt signiert werden, wohingegen es mit mehreren Clients im Durchschnitt unter 2% sind[128].

Dieser punktuelle Ausschnitt von der Verwendung von S/MIME betrifft nur einen sehr geschlossenen Kosmos, aber kann als eine obere Schranke betrachtet werden. Denn im Gegensatz zu anderen Organisationen und Nutzerinnen bietet die Universität über die DFN [3] eine funktionierende Infrastruktur, welche von gängigen Anwendungen unterstützt wird. Dies wird daran deutlich, dass alleine zwei Drittel der Zertifikate von der Deutschen Telekom ihr Vertrauen erben [128].

OpenPGP

S/MIME hat eine starke kommerzielle Komponente. Denn das Vertrauen in ein Zertifikat wird durch eine Kette von signierten Zertifikaten einer CA innerhalb der PKI gewährleistet. Viele CAs sind kommerzielle Anbieter und verlangen oftmals eine Gebühr. Eine Alternative dazu wird in den RFC3156 [114] und RFC4480 [78] mit OpenPGP spezifiziert. Mit OpenPGP können ebenso wie mit S/MIME E-Mails Ende-zu-Ende verschlüsselt werden oder mit digitalen Signaturen von der Nutzerinnen signiert werden. Obwohl sie auf algorithmischer Ebene ähnliche oder gleiche Verschlüsselungsalgorithmen und Signaturalgorithmen nutzen, sind sie auf der Protokollebene nicht kompatibel. Im Gegensatz zu S/MIME wird OpenPGP von einer Open-Source-Community entwickelt und gefördert. Der wesentliche Unterschied zu S/MIME ist die Herkunft von Vertrauen für einen öffentlichen Schlüssel bzw. für ein Zertifikat. Statt einem zentralistischen PKI-Ansatz wird ein dezentrales Netzwerk zur Vertrauensbildung bevorzugt. Personen können gegenseitig ihre öffentlichen Schlüssel überprüfen und nach einer erfolgreichen Verifikation können sie fremde öffentliche Schlüssel mit ihrem eigenen privaten Schlüssel signieren. Damit kann eine andere Personen nachvollziehen, wer diesem Schlüssel vertraut. Damit bildet sich ein dezentrales Netz von Schlüsseln, denen vertraut wird. Dieses Netz wird als Web of Trust bezeichnet.

Ein weiterer Unterschied ist, dass zum Schlüsselaustausch in der Regel ein dedizierter Server verwendet wird und von dort wird ein Schlüssel heruntergeladen. Dieser Server ist vergleichbar mit einem Telefonbuch. Alternativ kann ein Schlüssel auf einer Webseite oder auf anderen Plattformen zur Verfügung gestellt werden. Im Gegensatz dazu wird bei S/MIME das Zertifikat mit jeder Signatur mitgesendet. Dieser

[3] Deutsches Forschungsnetzwerk (DFN) nutzt als Rootzertifikat die Deutsche Telekom, welches von gängigen Anwendungen und Betriebssystemen standardmäßig als vertrauenswürdig eingestuft wird.

Ansatz wird mit Autocrypt[4] verfolgt. In diesem Ansatz kann ein öffentlicher Schlüssel in den Meta-Daten einer E-Mail mitgesendet werden.

Bei der praktischen Nutzung ist es, auffällig, dass viele gängige E-Mail-Anwendungen zwar standardmäßig S/MIME unterstützen, aber selten OpenPGP. Eine Ausnahme ist seit 2020 die E-Mail-Anwendung Thunderbird von Mozilla ab Version 78 [5]. Für andere Anwendungen muss meist ein Plugin installiert werden.

In der Studie von Stransky et al. wurden im Vergleich zu OpenPGP zwar knapp nur 6.000 mehr E-Mails mit S/MIME verschlüsselt, aber dafür etwa 1,75 Millionen mehr E-Mails mit S/MIME signiert [128]. OpenPGP wird somit deutlich weniger verwendet und dies wird ebenso bei der Verbreitung der öffentlichen Schlüsseln erkennbar. Es wurden etwa 9.800 S/MIME-Zertifikate gefunden, aber nur 3.700 OpenPGP Schlüssel [128]. Interessanterweise gibt es aber Personen, welche sowohl OpenPGP als auch S/MIME nutzen [128].

2.1.3 Zusammenfassung

Die Standardprotokolle bieten zunächst keinen Schutz der Vertraulichkeit, Integrität und Authentizität der E-Mail und deren Eco-System. Dies beginnt bei der fehlenden Authentifizierung der Nutzerinnen im RFC788 [101], welcher SMTP erstmalig spezifiziert und anfangs wurde eine verschlüsselte Übermittlung der E-Mail nicht berücksichtigt.

Erst durch Zusatzprotokolle, welche in den Jahren immer wieder neu hinzugekommen sind und immer wieder angepasst werden, kann die Sicherheit erhöht werden. Die Tabelle 2.1 fasst die Sicherheitseigenschaften zusammen. Neben diesen spezifizierten Protokollen gibt es noch eine Reihe von anderen Sicherheitsprotokollen auf Basis der E-Mail, wie zum Beispiel die DE-Mail[6], Tutanota[7], Volksverschlüsselung[8] oder Besondere elektronische Anwaltspostfach[9]. Diese sind als RFC nicht spezifiziert und darum wird auf eine Betrachtung verzichtet.

Bei den RFC-Protokollen muss vor allem abgewogen werden, ob dem Server vertraut werden muss oder nicht. Dieses Vertrauen bezieht sich nicht nur darauf, mit wem die Daten nach der Gesetzgebung geteilt werden müssen, sondern wie gut sie dort geschützt sind. Serverseitige Schutzmechanismen, wie TLS, SPF und DKIM, sind mittlerweile weit verbreitet. Doch die Umsetzung von strikten Regeln und das Verwerfen von Nachrichten ist nicht sehr stringent und eine Ursache dafür ist sicherlich die hohe Fehlerquote bei der Nutzung von DKIM. E-Mail-Provider tendieren anscheinend dazu, Nutzerinnen lieber zu viele als zu wenige Nachrichten zuzustellen, und entscheiden sich im Zweifelsfall lieber für die Verfügbarkeit einer Nachricht als für die Sicherheit. Diese Konzepte sind zunächst ohne die Interaktion zwischen Mensch und

[4] https://autocrypt.org/level1.html, letzter Zugriff 4. Mai 2023, 18:00
[5] https://www.thunderbird.net/de/thunderbird/78.0/whatsnew/?locale=de&version=78.2.2&channel=release&os=%25OS%25&buildid=%25APPBUILDID%25&oldversion=68.12.0
[6] https://www.de-mail.info/
[7] https://tutanota.com/
[8] https://volksverschluesselung.de/
[9] https://www.bea-brak.de/bea/index.xhtml?dswid=-1045

Verfahren	Vertraulichkeit	Integrität	Authentizität	Nichtabstreitbarkeit
TLS	◑	◑		
SPF			○	
DKIM		○	○	○
S/MIME	●	●	●	●
OpenPGP	●	●	●	●

Tabelle 2.1: ● bedeutet, dass die Sicherheitseigenschaft zwischen beiden Endgeräten sichergestellt wird. ◑ bedeutet, dass die Sicherheitseigenschaft nur zwischen Server und Server sowie Endgerät und Server sichergestellt wird. ○ bedeutet, dass die Sicherheitseigenschaft nur zwischen Server und Server sichergestellt wird.

Computer entwickelt worden. Demgegenüber steht die Verwendung von Ende-zu-Ende-Verschlüsselung und deren digitalen Signaturen. Diese erfordern menschliche Interaktionen und werden direkt mit der Anwendung der Nutzerinnen umgesetzt. Die Anwendung von diesen Sicherheitsmechanismen an einer Universität deutet darauf hin, dass diese wahrscheinlich nicht weit verbreitet sind und in der Breite keine Anwendung finden. Nur Subgruppen mit einem besonderen Bedürfnis an Ende-zu-Ende-Verschlüsselung nutzen diese Sicherheitsmechanismen. In der Konsequenz muss aus einer technischen Perspektive zwar zwischen verschiedenen Sicherheitsmechanismen und damit zwischen unterschiedlich sicheren E-Mails unterschieden werden. Die meisten Nutzerinnen werden damit aber eher selten konfrontiert und unterscheiden damit nicht zwischen den unterschiedlich sicheren E-Mails. Gleichzeitig bietet die E-Mail sehr viele Möglichkeiten zur Überprüfung der Sicherheit, welche Nutzerinnen bei der Entscheidung helfen.

Zusätzlich ist durch diese vielen Ergänzungen und zusätzlichen Protokolle ein komplexes und teilweise sehr fragiles System entstanden. In Bezug auf die Sicherheit sind dadurch aber eine Vielzahl an möglichen Angriffsvektoren entstanden. Dies betrifft die Kompatibilität zu vorherigen und veralteten Spezifikationen sowie durch die Kombination von verschiedenen unabhängigen Varianten [92, 93, 100].

2.2 Nutzung von E-Mails

Die E-Mail wurde oft totgesagt, aber ist immer noch ein wesentlicher Bestandteil im Alltag vieler Menschen und die Interaktion von Menschen mit E-Mails wird seit Jahrzehnten erforscht. Hierbei wird deutlich, dass die Nutzung von E-Mails sich in der Zeit gewandelt hat, aber ein sehr allgemeines Werkzeug ist.

In einer Nutzerstudie ($N = 20$) von 1996 untersuchten Whittaker und Sidner Meta-Daten über die E-Mail-Nutzung der Personen und führten jeweils Interviews [145]. In den Interviews wird wiederholt betont, dass die E-Mail nicht nur zur Kommunikation verwendet wird, sondern auch für andere Aufgaben, wie ein persönliches Archiv oder Aufgabenverwaltung. Diese zusätzlichen Verwendungszwecke fassen Whittaker und Sidner unter dem Begriff *E-Mail Overload* zusammen. Gleichzeitig können sie beobachten, dass der Posteingang durchschnittlich 1.624 E-Mails umfasst und täglich durchschnittlich 49 E-Mails neu eingehen. Eine Übersicht und ein Vergleich mit späteren Messungen ist in Tabelle 2.2 darge-

Jahr	1996	2006	2012	2012
Veröffentlichung	[145]	[45]	[53]	[53]
Kontext	Arbeit	Arbeit	Arbeit	Privat
# Nutzerinnen	18	600	17	19
Anwendung	NotesMail	Outlook	Gmail	Gmail
# Erhaltene E-Mails pro Tag	49	87 (59)	NA	NA
Größe vom Posteingang	1.624	1.150 (512)	3.003 (1.483)	15.030 (3.500)
# Ungelesene im Posteingang	NA	153 (7)	696 (3)	4.846 (421)
# Ordner bzw. Labels	47	133 (77)	27 (9)	22 (11)

Tabelle 2.2: Das durchschnittliche Nutzungsverhalten im Kontext der E-Mail aus unterschiedlichen Studien wird dargestellt. In Klammern wird jeweils der Median dargestellt. Ähnliche Tabellen sind in den Arbeiten von Fisher et al. und Grevet et al. enthalten [45, 53]. Bei Gmail wurden statt Ordner Labels gemessen.

stellt. In den Interviews berichten die Nutzerinnen neben positiven Aspekten auch über negative Folgen dieser Übernutzung der E-Mail. Zehn Jahre später untersuchten Fisher et al. die E-Mail-Postfächer von 600 Personen [45]. Die durchschnittliche Posteingangsgröße war im Durchschnitt ähnlich, aber in der Gesamtanzahl an E-Mails pro Person gibt es deutliche Unterschiede. Durchschnittlich hatte eine Person in der Studie von 1996 2.482 E-Mails und in der Studie von 2006 wurden 28.660 E-Mails pro Nutzer gezählt. Ein weiterer Unterschied ist, dass in der Studie von 2006 die Personen im Durchschnitt 86 E-Mails erhalten haben. Fisher et al. berechnen jeweils sowohl den Durchschnittswerte als auch den Median, und in Tabelle 2.2 werden diese dargestellt. Es gibt große Unterschiede zwischen dem Mittelwert und dem Median. Folglich gibt es sehr wahrscheinlich große Unterschiede zwischen einzelnen Personen. Beide Studien wurden im beruflichen Kontext durchgeführt, aber in unterschiedlichen Unternehmen mit unterschiedlich verwendeten Anwendungen. Eine kontinuierliche Entwicklung kann daraus nicht geschlussfolgert werden, aber Fisher et al. zeigen, dass die Bewältigung einer E-Mail-Flut im beruflichen Kontext weiterhin eine Herausforderung ist.

Grevet et al. interviewten und untersuchten die Postfächer von 19 Personen [53]. Sie konnten dabei oftmals das berufliche als auch das private Postfach untersuchen, wobei alle Personen Gmail nutzten. Im Gegensatz zu vorherigen Studien hatten ihre Personen im Durchschnitt 3.003 E-Mails im beruflichen Posteingang und damit hat sich die Anzahl der E-Mails mehr als verdoppelt. Diese Veränderung kann unterschiedliche Ursachen haben, wie die Autorinnen selbst anmerken. Als eine Möglichkeit nennen sie insbesondere die vielen Funktionen von Gmail, wie eine verbesserte Suche oder das Markieren von E-Mails [53]. Auffällig ist aber der Unterschied zwischen den beruflichen und persönlichen Postfächern der gleichen Personen. Die persönlichen Posteingänge beinhalten im Durchschnitt 15.030 E-Mails und damit fünfmal so viele im Vergleich zum beruflichen Posteingang. Gleichzeitig ist auffällig, dass persönliche Posteingänge deutlich mehr ungelesene E-Mails im Durchschnitt beinhalten.

2.2.1 Private Kommunikation

Eine Studie von Yahoo bestätigt den Eindruck. Castro et al. untersuchten das Verhalten von Nutzerinnen von persönlichen E-Mail-Postfächern [40]. 90% der empfangenen E-Mails waren automatisch, also von Computern erzeugt. Dies sind beispielsweise Bestellbestätigungen oder Versandbenachrichtungen. Viele dieses E-Mails verbleiben im Posteingang und werden weder gelesen noch verschoben oder gelöscht. Castro et al. unterteilen die Nutzerinnen in unterschiedliche Aktivitätsgruppen. In diesen Gruppen wurden 15,3% bis 20,2% der E-Mails im Posteingang gelesen und stattdessen wurden teilweise bis fast 20% der E-Mails ohne zu lesen gelöscht [40]. Gleichzeitig wurden nur 2% aller eingehenden E-Mails beantwortet.

Folglich ist es im privaten Kontext durchaus üblich, nicht alle E-Mails zu bearbeiten und viele zu ignorieren oder ungelesen zu löschen. In diesem Kontext ist es bemerkenswert, dass Phishing-E-Mails überhaupt geöffnet werden und diese hohe Hürde anscheinend überwinden können. In einem erfolgreichen Phishing-Angriff mittels E-Mail gehört diese E-Mail somit zu den relevanten E-Mails, welche überhaupt beachtet werden. Die hohe Ignoranzrate von E-Mails kann bei Feldstudien zum Thema Phishing das Ergebnis verfälschen. Denn eine Möglichkeit ist, dass Teilnehmerinnen in einer Studie die eingegangen Phishing-E-Mail gar nicht wahrgenommen haben oder als nicht relevant eingestuft haben. Folglich haben sie dann den Stimulus ignoriert und dies schränkt die Aussagekraft der Studie weiter ein.

In einer Umfrage von Bentley et al. wurden Nutzerinnen von Yahoo nach ihrer E-Mail-Nutzung befragt. Für 41% der Teilnehmerinnen ist die E-Mail nicht wichtig für eine persönliche Kommunikation, aber die Mehrheit empfängt kommerzielle E-Mails, wie zum Beispiel Rechnungen, Bestellinformationen, Reiseinformationen oder Gutscheine [14]. Dies ist ein großer Unterschied zu Messengern wie WhatsApp, Signal oder Threema, welche hauptsächlich zur persönlichen Kommunikation verwendet werden.

Alrashed et al. untersuchten die Nutzungsaktivitäten, wie zum Beispiel Löschen, Markieren, Überfliegen (kurze Lesezeit), Lesen, Antworten, Weiterleiten von E-Mails oder Link-Klicks in E-Mails, von etwa 500 Millionen privaten Postfächern mit etwa 17 Milliarden E-Mails und unterscheiden dabei zwischen implizit signifikanten E-Mails und implizit nicht signifikanten E-Mails [5]. E-Mails sind implizit signifikant, wenn E-Mails von Nutzerinnen in den Fokus-Tab in Outlook bewegt werden. E-Mails sind implizit nicht signifikant, wenn die E-Mails von Nutzerinnen vom Fokus-Tab in Outlook weg bewegt werden. Hierbei ist anzumerken, dass Outlook ermöglicht, E-Mails automatisch in einen Fokus-Posteingang für wichtige E-Mails und einen anderen Posteingang vorzusortieren. Bei implizierten nicht signifikanten E-Mails waren nur in etwa 10% der Aktivitäten das Markieren von E-Mails, Öffnen eines Links bzw. eines Anhangs. Alrashed et al. geben für diese Aktivitäten keine einzelnen Prozentangaben an. Bei implizit signifikanten E-Mails waren 14,68% der Aktivitäten ein Link-Klick und in 6,51% das Öffnen einer E-Mail. Demgegenüber sind die Löschung (33,5%) und die Verschiebung in einen anderen Ordner (30,6%) die häufigsten Aktivitäten bei implizit nicht signifikanten E-Mails. Bei implizit signifikanten E-Mails sind längeres Lesen (24,82%) und Löschung (21,97%) die häufigsten Aktivitäten. Das Klicken auf Links und Anhänge ist bei manchen E-Mails (implizit signifikanten E-Mails) üblich und erfolgt regelmäßig. Folglich ist es bei Phishing-E-Mails plausibel, dass diese Aktionen getätigt werden.

2.2.2 Berufliche Kommunikation

Alrashed et al. untersuchten neben persönlichen Postfächern bei 170 Millionen beruflichen Postfächern die Nutzungsaktivitäten. Für das Öffnen von Links und Anhängen gibt es weder bei implizit signifikanten noch bei nicht signifikanten E-Mails eine Prozentangabe, aber in der Summe mit Markierungen als Spam/ungelesen, Antworten und Weiterleitungen sind es weniger als 11,5% aller Aktivitäten [5]. Im beruflichen Kontext dominiert das Lesen von E-Mails. 69,2% bei implizit nicht signifikanten E-Mails und 75,83% bei implizit signifikanten E-Mails. Folglich ist es im beruflichen Kontext eher unüblich, auf Links zu klicken oder Anhänge zu öffnen. Dies bietet damit eine Möglichkeit, Personen besser vor gefährlichen Links oder Anhängen zu schützen.

2.2.3 Zusammenfassung

Die E-Mail ist immer noch ein alltägliches Kommunikationsmittel und wird sowohl im beruflichen als auch im privaten Kontext viel benutzt. Es gibt aber deutliche Unterschiede zwischen diesen Kontexten. In beiden Kontexten erhalten Personen viele E-Mails pro Tag und werden teilweise damit überflutet. Damit werden Fehler wahrscheinlicher.

Die Nutzung unterscheidet sich zwischen beruflichen und privaten E-Mails. Vor allem im privaten Kontext werden viele E-Mails maschinell erzeugt und werden nicht beachtet. Im Kontext von Phishing-E-Mails scheint es darum erstrebenswert, dieses Verhalten bei Personen zu unterstützen und Phishing-E-Mails nicht zu beachten. Damit wird der Angriff verhindert. Bei einem Phishing-Angriff wird häufig ein Link geöffnet und dies ist in der privaten Nutzung alltäglich und häufige Aktivität. Darum ist es erstrebenswert, Personen im Kontext von Phishing-E-Mails davon abzuhalten und so Fehler zu verhindern.

Das Nutzungsverhalten im Kontext von E-Mails verdeutlicht nochmals, dass Ende-zu-Ende-Verschlüsselung nur selten bedeutsam ist. Im privaten Kontext ist dies oftmals unnötig, denn bei Maschinen-erzeugten E-Mails oder E-Mails von (kommerziellen) Organisationen kann die Absenderseite leicht zur Kooperation mit staatlichen Stellen verpflichtet und gezwungen werden und andere E-Mails sind im Allgemeinen eher selten im privaten Kontext.

Im beruflichen Kontext hat die Organisation in der Regel sowieso ein berechtigtes Interesse oder die Pflicht, in bestimmten Fällen die Kommunikation nachvollziehen zu können und dies ohne kooperatives Verhalten der betroffenen Personen. Eine Organisation hat darum im Allgemeinen gar kein Interesse an einer echten Ende-zu-Ende-Verschlüsselung, wenn nicht andere Maßnahmen noch getroffen werden. Es gibt dazu Ausnahmen bei der Nutzung von Ende-zu-Ende-Verschlüsselung. Dies betrifft zum Beispiel Journalistinnen und Aktivistinnen; Studien zeigen in diesen Fällen die Nutzung von Ende-zu-Ende-Verschlüsselung [21, 85, 81].

Kapitel 3

Die Kunst vom Angeln

Im vorherigen Kapitel wurde die E-Mail mit ihren verschiedenen Spezifikationen und deren Verwendung vorgestellt. In diesem Kontext stellen Phishing-Angriffe eine besondere Gefahr dar und im folgenden Kapitel werden diese Angriffe vorgestellt. Es wird zunächst ein informelles Gefahrenmodell entwickelt und mit exemplarischen Angriffen aus der Praxis verdeutlicht. Das Gefahrenmodell wird dann im folgenden Kapitel formalisiert.

3.1 Gefahrenmodell

Die NIST[1] definiert Phishing aufbauend auf dem RFC4949 [122] wie folgt:

> A technique for attempting to acquire sensitive data, such as bank account numbers, through a fraudulent solicitation in email or on a web site, in which the perpetrator masquerades as a legitimate business or reputable person.

Diese Definition beinhaltet mehrere unterschiedliche Aspekte. Zunächst beginnt die Definition mit einem konkreten Ziel der Angreiferin. Die Angreiferin möchte Zugriff auf sensitive Daten erhalten. Dies können Zugangsdaten zu verschiedenen Benutzerkonten sein. Diese Zugangsdaten oder andere sensitive Informationen werden durch eine Person, der Nutzerin, freigegeben. Um dieses Ziel zu erreichen, täuscht die Angreiferin vor, eine Person oder Organisation mit einem legitimen Interesse zu sein.

Bei diesem Angriff sind also drei Parteien beteiligt: Die Angreiferin, Nutzerinnen (also die Zielpersonen des Angriffs) und eine (legitime, vertrauenswürdige) Dritte-Partei, wobei diese nicht zwangsläufig aktiv ist. Bei dieser Täuschung ist hervorzuheben, dass die Nutzerinnen entscheiden, die sensitiven Informationen preiszugeben. Im Vordergrund steht die Täuschung des Menschen und damit sind diese Angriffe

[1] Das *National Institute of Standards and Technology* (NIST) ist eine US-Bundesbehörde zur Standardisierung. Unter anderem wurden durch diese Behörde kryptographische Algorithmen standardisiert.

nicht zwangsläufig durch die typischen kryptographischen Techniken, wie die Authentizität einer Nachricht, zu verhindern. Dhamija et al. benennen fehlendes Wissen, visuelle Täuschungen und begrenzte Aufmerksamkeit als mögliche Ursachen für eine Täuschung [38]. Angreiferinnen nutzen Schwächen in der Kognition oder Psyche der Nutzerinnen für eine erfolgreiche Täuschung aus [134]. Die Angriffskanäle sind vielseitig und umfassen Kommunikationskanäle, wie zum Beispiel E-Mail, Web, SMS, Messengers sowie soziale Netzwerke, wie Facebook oder Twitter. Im Folgenden werden Angriffe per E-Mail betrachtet. Diese haben oftmals eine Fortsetzung im Web-Browser. Denn die Freigabe von Zugangsdaten erfolgt meist durch Eingabe eines Passworts auf einer Webseite.

Phishing-Angriffe

Ein Phishing Angriff besteht aus zwei Aspekten:

1. Die Angreiferin täuscht eine legitime Partei gegenüber den Nutzerinnen vor.
2. Die Nutzerinnen tätigen eine Aktion, welche nur durch Aufforderung durch die legitime Partei erfolgt.

3.2 Fallbeispiele

In diesem Abschnitt werden einige kurze Fallbeispiele skizziert und die Vielseitigkeit der tatsächlichen Angriffe dargestellt. Phishing-Angriffe können sich gegen einzelne Personen richten oder gegen einen größeren Personenkreis. Die Auswirkungen davon sind, dass diese Angriffe unterschiedlicher Qualität sind und unterschiedliche Gegenmaßnahmen ermöglichen.

3.2.1 Massenangriffe

Viele Privatpersonen finden in ihrem E-Mail-Postfach regelmäßig Phishing-Nachrichten von Online-Bezahldiensten, Banken oder Lieferdiensten.

Abbildung 3.1 zeigt verschiedene Angriffe per E-Mail. Im ersten Beispiel wird der Paketzusteller DHL als Herkunft der Nachricht vorgetäuscht. Diese E-Mail wirkt relativ einfältig. Die Anrede als Kunde mit E-Mail-Adresse ist nicht sehr persönlich und ein angebliches Paket mit einer zufälligen Trackingnummer wird als Grund für den Besuch einer Webseite genannt. Die E-Mail zeigt zwar ein Logo der Post an, jedoch weicht die Darstellung von bekannten E-Mails von DHL ab. Die Erkennung dieser E-Mail als eine Phishing-Nachricht ist darum relativ einfach und der E-Mail-Provider hat diese bereits als Spam bzw. Phishing markiert.

Die zweite E-Mail soll vom Zoll kommen und es werden Probleme mit einem Paket beim Zoll suggeriert. Die Nutzerin soll als Aktion zur Lösung einen Geldbetrag an eine Webseite bezahlen. Diese E-Mail ist weiterhin nicht personalisiert, aber der Text wirkt deutlich professioneller. Diese E-Mail wurde vom E-Mail-Provider als Spam bzw. Phishing erkannt.

Im dritten Beispiel stammt die E-Mail angeblich von Amazon und täuscht eine Warnung vor einer ungewöhnlichen Anmeldung vor. Im Gegensatz zu den vorherigen Angriffen ist diese bereits personalisiert und der Provider hat diese nicht als Phishing-Angriff erkannt. Ziel des Angriffes ist es, die Webseite der Angreiferin einzugeben. Die Webseite ähnelt der validen Amazon-Webseite und es wird die Eingabe von den Zugangsdaten verlangt. Bei dieser Täuschung gibt die Angreiferin vor, dass die E-Mail und die Webseite von Amazon ist. Diese Anfrage nach den Zugangsdaten ist unter dieser Annahme legitim und vertrauenswürdig. Einem beliebigen Dritten würden die Zugangsdaten verweigert werden.

Der Besuch der Webseite der Angreiferin ist das Ziel der E-Mail. Wenn dieses (Zwischen)-Ziel erreicht wurde, ist das nächste Ziel die Eingabe der Zugangsdaten und erst dann wurde das Hauptziel des Angriffes erreicht.

Diese Angriffe, auch der personalisierte, erfolgen in der Regel gegen viele Menschen. Die Angreiferin ist dabei opportunistisch gegenüber den vielen Nutzerinnen und bereits wenige erfolgreiche Angriffe sind dabei ausreichend. Die Gegenmaßnahmen der E-Mail-Provider ist dabei die Blockade von diesen Angriffen, nachdem diese aufgedeckt wurden. Für die einzelne Person kann der Schadensfall dennoch erheblich sein. Die Blockierung von Webseiten im Browser ist ebenso eine gängige und erforschte Methode. Google Chrome, Mozilla Firefox und Safari von Apple nutzen dabei dieselbe Blockliste und diese Liste schützt die Mehrheit der Browsernutzerinnnen [97]. Oest et al. haben ein Testframework entwickelt, um verschiedene Blocklisten von Phishing-Seiten zu untersuchen [97]. Nach ihren Ergebnissen wird eine Phishing-Webseite innerhalb von wenigen Stunden nach Meldung blockiert. Dies verringert das Angriffsfenster für die obigen Phishing-Angriffe.

3.2.2 Gezielte Angriffe

Eine andere Alternative sind gezielte Angriffe gegen eine bestimmte Person oder einen kleinen Personenkreis. Die Personenkreise können dabei sehr unterschiedlich sein. Warford et al. haben ein Framework für unterschiedliche *at-risk users* konstruiert und insbesondere Journalistinnen, Aktivistinnen, Mitarbeiterinnen in Nicht-Regierungsorganisationen (NGO), Personen aus der Politik (inklusive Mitarbeiter in den Büros von Politikerinnen) wurden als Gruppen mit Zugriff auf sensible oder vertrauliche Informationen identifiziert [140]. Diese Personengruppen sind besonders Phishing-Angriffen ausgesetzt, denn mittels einem erfolgreichen Phishing-Angriff können Zugangsdaten und damit der Zugriff auf vertrauliche Dokumente erlangt werden. Ein bekanntes Beispiel für diese Angriffe gegen diese Personen ist ein Angriff auf Billy Rinehart, einen Wahlkampfmanager der Demokraten im US-Wahlkampf 2016. Abbildung 3.2 zeigt die Darstellung vom Inhalt einer E-Mail an ihn. In dieser E-Mail wurde eine Meldung von Google über einen ungewöhnlichen Anmeldeversuch auf seinen Gmail-Account vorgetäuscht und er wurde aufgefordert eine Webseite zu besuchen, um sein Passwort zu ändern. Nach Rücksprache mit seinem IT-Support gab es ein Missverständnis und in der Folge wurden seine Zugangsdaten gestohlen. Diese Zugangsdaten wurden genutzt, um interne E-Mails zu veröffentlichen und es wurde dadurch ver-

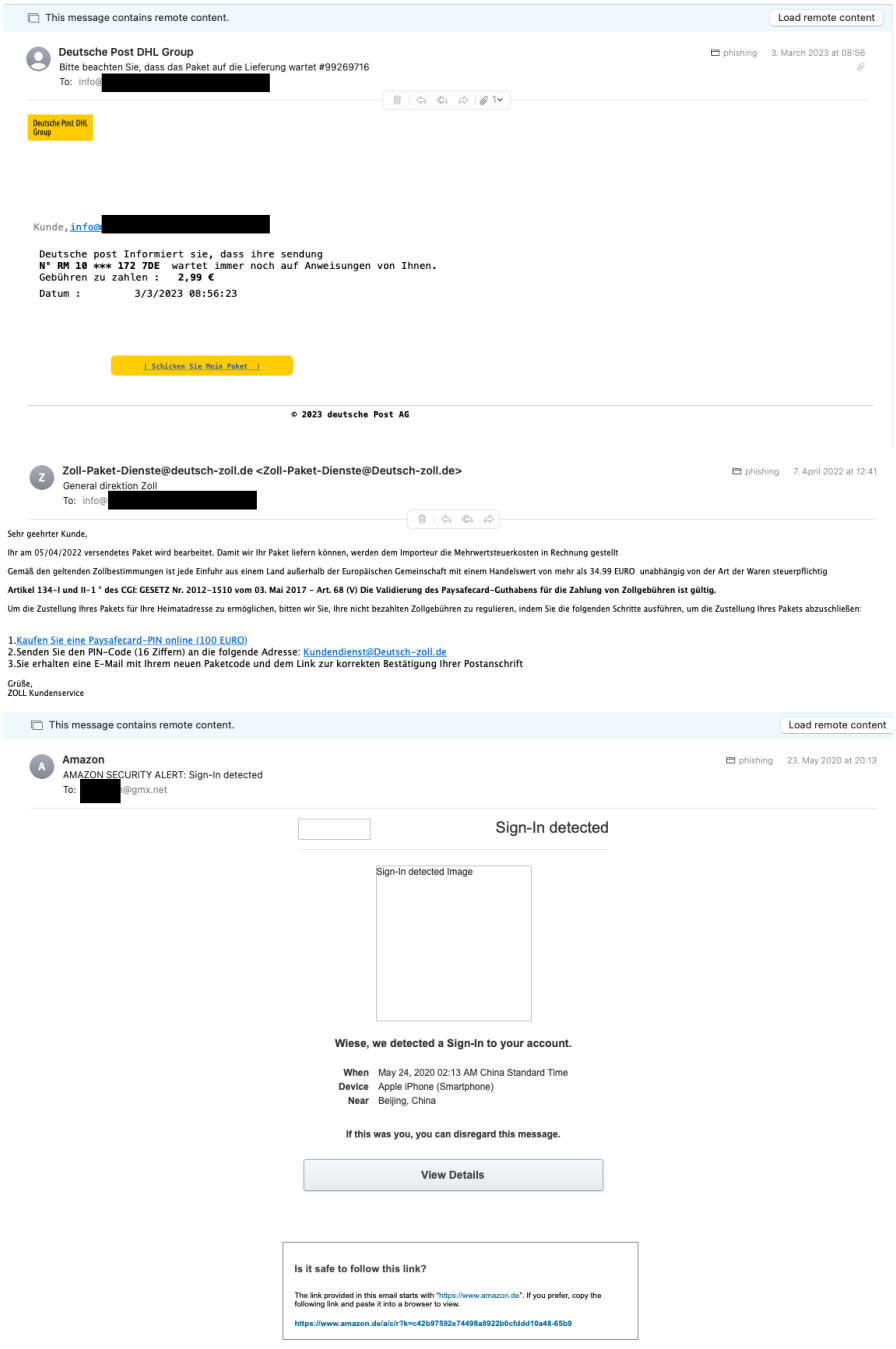

Abbildung 3.1: Sammlung von verschiedenen Angriffen unterschiedlicher Qualität.

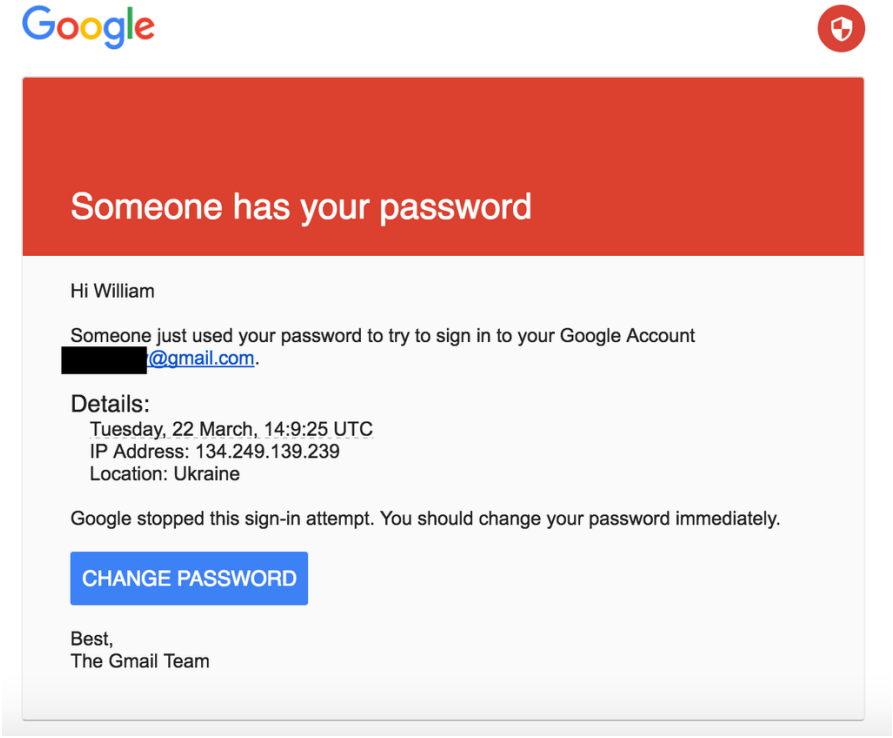

Abbildung 3.2: Diese E-Mail hat Billy Rinehart im Postfach gefunden.

sucht, den US-Wahlkampf zu beeinflussen. [2] Dieser Phishing-Angriff und weitere im US-Wahlkampf 2016 verdeutlichen die extremen Gefahren und Auswirkungen von Phishing-Angriffen.

Das Citizenlab der Universität Toronto hat Phishing-Angriffe gegen Mitarbeiterinnen von NGOs als Fallstudien gesammelt und veröffentlicht. Beispielsweise wurden Phishing-Angriffe gegen ägyptische Bürgerrechtsbewegungen aufgezeigt[3]. Von Ende November bis Ende Dezember 2016 konnten sie 92 Angriffe gegen Bürgerrechtlerinnen, Journalistinnen und Anwältinnen identifizieren. Alle diese Personen waren beim gleichen politischen Thema aktiv und die Angriffsnachricht nahm darauf Bezug. Die verwendete Server und das gemeinsam verwendete Phishing-Toolkit[4] ermöglicht die Verknüpfung der verschiedenen Angriffen als eine Angriffsserie. In dem Zeitraum wurde ein ägyptischer Anwalt verhaftet und in dieser Phishing-Kampagne wurde in der Nachricht an die Nutzerinnen bereits kurz nach der Festnahme auf eben diese Bezug genommen und auf eine PDF in einer Dropbox verwiesen. Abbildung 3.3

[2] Die genauen Hintergründe wurden von der New York Times im Dezember 2016 veröffentlicht. URL zum Artikel: https://www.nytimes.com/2016/12/13/us/politics/russia-hack-election-dnc.html
Der Guardian hat diesen Bericht aufgegriffen: https://www.theguardian.com/us-news/2016/dec/14/dnc-hillary-clinton-emails-hacked-russia-aide-typo-investigation-finds

[3] Der Nilephish-Bericht ist unter https://citizenlab.ca/2017/02/nilephish-report/ online verfügbar. Letzter Zugriff 23.3.23 um 16:00

[4] Ein Phishing-Toolkit ermöglicht das einfache Erstellen der Infrastruktur, wie zum Beispiel E-Mails, Webseiten für einen Angriff. Diese sind frei verfügbar oder können käuflich erworben werden [17].

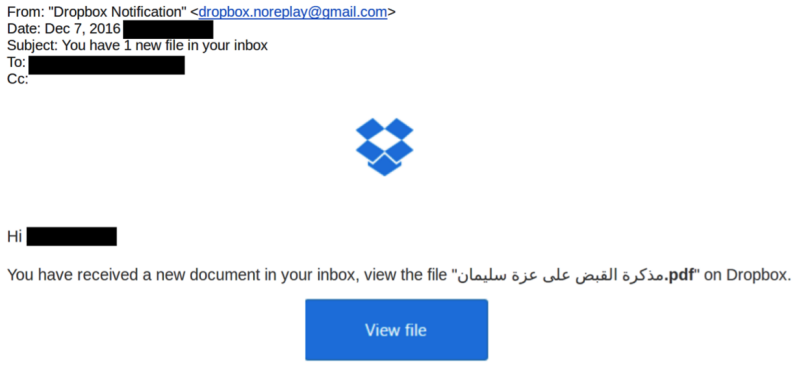

Abbildung 3.3: Phishing E-Mail

stellt eine solche Phishing-Nachricht dar. Der Link in der E-Mail verwies wieder auf eine gefälschte Zugangswebseite von Dropbox. Diese ist in Abbildung 3.4 dargestellt.

In einer anderen Phishing-Kampagne wurden tibetische Aktivistinnen gezielt angegriffen.[5] Diese Angriffe liefen über ein Jahr und die Domains wechselten öfter. Gleichzeitig wurden nur wenige Angriffsnachrichten gesammelt. Bei diesen gezielten Angriffen sind die Blocklisten wahrscheinlich weniger effizient, weil diese E-Mails erst gemeldet bzw. erkannt werden müssen und danach erst blockiert werden. Insbesondere die Kampagne gegen die tibetischen Aktivistinnen mit den Änderungen der Angriffsdomains sind damit gefährlich. Gleichzeitig können innerhalb der kurzen Zeit bis zur Blockierung immer noch Nutzerinnen erfolgreich angegriffen werden und der Schaden kann beachtlich sein.

3.2.3 Gemeinsamkeiten

Die obigen Beispiele unterscheiden sich zwar in der Qualität des Angriffs, aber es gibt einige Gemeinsamkeiten. Die Angriffe hatten folgende Eigenschaften:

1. Der Angriff erfolgt in **mehreren Stufen**. Wenn Nutzerinnen die E-Mail ignoriert und den Link nicht öffnen, dann ist der Angriff nicht erfolgreich.
2. Der Erfolg des Angriffs ist im Wesentlichen von **Entscheidungen der Nutzerinnen** abhängig. Wenn die Nutzerinnen sich entscheiden, die E-Mail zu ignorieren oder das Passwort nicht eingeben, dann ist der Angriff nicht erfolgreich.
3. Bei diesen Angriffen muss nicht (nur) die Anwendung getäuscht werden, sondern es erfolgt eine **Täuschung des Menschen**.

Allerdings sind die Ziele der Angreiferinnen unterschiedlich. Die Ziele reichen von finanziellen Interessen über Spionage gegen die Zivilgesellschaft bis zu (außen)-politischen Interessen. Damit sind die

[5] Der Bericht ist unter https://citizenlab.ca/2018/01/spying-on-a-budget-inside-a-phishing-operation-with-target online verfügbar.

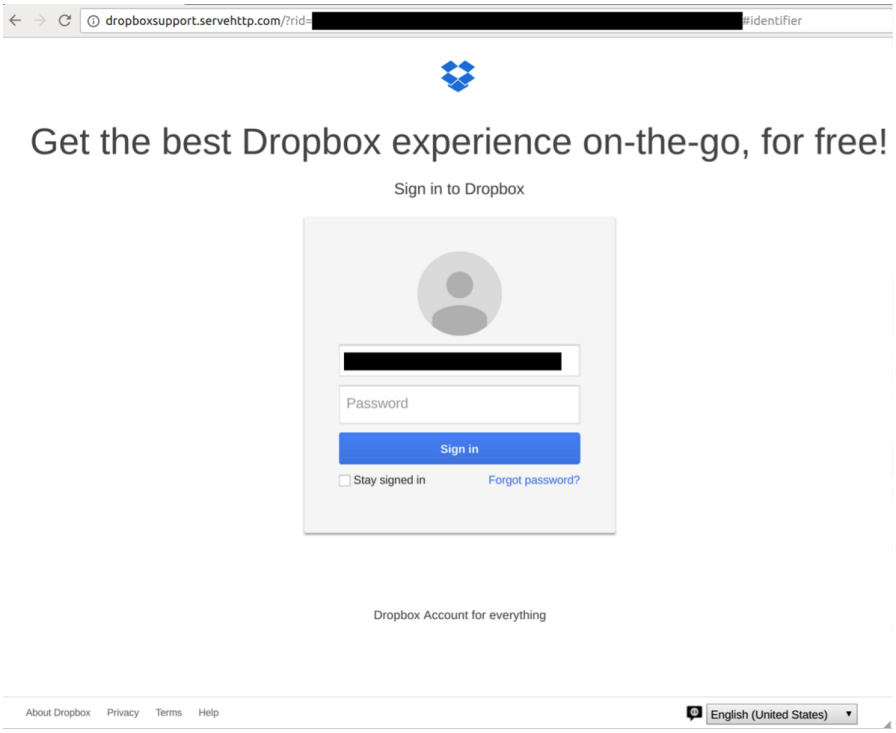

Abbildung 3.4: Phishing-Webseite

Akteure mit ihren Fähigkeiten sehr unterschiedlich. Der Verdacht liegt nahe, dass sogar staatliche Akteure Phishing-Angriffe einsetzen. Die Gegenmaßnahmen von Servern zielen häufig auf die Erkennung und Verhinderung von größeren Phishing-Kampagnen ab und den Schutz der Mehrheit der Nutzerinnen, wobei die einzelne Nutzerin nicht immer geschützt werden kann. Damit müssen die Anwendungen der Nutzerinnen Gegenmaßnahmen ergreifen, um den individuellen Schutz zu verbessern.

Zielgruppen von Phishing

Phishing-Angriffe können sich gegen große Gruppen, wie zum Beispiel potentielle Kundinnen von PayPal, DHL oder Amazon, richten, aber die Zielgruppe kann sehr klein werden bis zu einzelnen Subgruppen, wie bestimmte Journalistinnen oder Politikerinnen.

3.3 Abgrenzung zu anderen Angriffen

Neben dem bereits vorgestellten Angriff gibt es noch weitere verwandte und ähnliche Angriffe. Statt dem Zugriff auf Nutzerkonten sind weitere Ziele:

1. Geldüberweisung
2. Ransomware-Verbreitung
3. Zugriff auf IT-Systeme
4. Zugriff auf Dokumente
5. Weiterleitung einer Nachricht

Die Angreiferin kann die Nutzerin bitten, eine Überweisung auf ein Konto zu tätigen. In diesem Szenario ist das klassische Beispiel, dass die Angreiferin vorgibt, die Chefin zu sein und die Nutzerin als Untergebene anweist, eine Überweisung zu tätigen. In diesem Fall ist die Täuschung über eine legitime dritte Partei naheliegend und diese ist der Nutzerin bekannt. Ein anderes klassisches Beispiel sind sogenannte Nachrichten von angeblichen Prinzen oder anderen wohlhabenden Personen, welche die Nutzerin bitten, einen kleineren Betrag zu überweisen, um ein größeres Vermögen zu erhalten. In diesen Fällen ist die Identitätstäuschung eher nebensächlich, denn der Prinz oder die wohlhabende Person ist meist erfunden und der Nutzerin unbekannt. Gleichzeitig wird diese Nachricht an Millionen Menschen versendet und damit kann dies als eine (bösartige) Spam-Nachricht[6] betrachtet werden. In solchen Randfällen ist die Trennung zwischen Spam und Phishing nicht mehr ganz eindeutig und fließend.

In den anderen Fällen sind die Angriffstechniken und die Täuschung der Nutzerinnen ähnlich. Oftmals erfolgt der Angriff in mehreren Stufen und die Nutzerin trifft die wesentlichen Entscheidungen um den Angriff zu einem Erfolg zu führen. Das Ziel ist ein anderes und die nötigen Schritte sind anders.

[6] Die NIST definiert Spam wie folgt: *Electronic junk mail or the abuse of electronic messaging systems to indiscriminately send unsolicited bulk messages.* [98]

Übergang zu anderen Angriffen

Der Übergang zu anderen Angriffsformen ist fließend und nicht immer eindeutig zu trennen.

Phishing-Angriffe haben sich zu einem allgewärtigen und vermeintlich einfachen Einfallstor für Ransomware und andere Angriffe auf IT-Systeme entwickelt. Nach einem erfolgreichen Phishing-Angriff werden beispielsweise Daten von einem Unternehmen verschlüsselt oder gestohlen, um im Anschluss ein Lösegeld zu erpressen.

3.4 Analysen zu Phishing in freier Wildbahn

Die obigen Fallbeispiele illustrieren, wie ein Phishing-Angriff ablaufen kann und wie diese Nachrichten und Webseiten aussehen. In diesem Abschnitt wird die wissenschaftliche Literatur zur Untersuchung von tatsächlichen Phishing-Angriffen vorgestellt.

Phishing-Angriffe können aus der Perspektive von E-Mail-Providern, welche Phishing-Nachrichten an ihre Kundinnen erhalten, und aus der Perspektive von Organisationen, welche als legitime Partei zur Täuschung der Nutzerinnen eingesetzt werden, betrachtet werden. Simoiu et al. haben zwischen dem 7. April und dem 31. August 2020 die eingehenden Phishing-Nachrichten unter den Gmail-Kundinnen gemessen [123]. Insgesamt haben sie etwa mehr als eine halbe Milliarde E-Mails, welche einen Link zu einer als Phishing markierten Webseite enthalten, identifiziert. Durchschnittlich waren wöchentlich 17 Millionen Menschen davon betroffen. Innerhalb der fünf Monate haben sie 400.000 verschiedene Phishing-Kampagnen beobachtet, wobei 91% der Kampagnen weniger als 1.000 E-Mails versendet haben. Allerdings gibt es sehr wenige Kampagnen, welche für einen Großteil der E-Mails verantwortlich sind. 80% der Kampagnen waren bereits nach einer Woche nicht mehr aktiv [123].

Van der Heijden und Allodi haben Phishing-Angriffe, welche ein großes europäisches Finanzunternehmen mit einem Milliarden-Umsatz und Millionen an Kundinnen nachahmten, untersucht [134]. Die Kundinnen des Unternehmens waren aufgefordert, Phishing-Angriffe, welche sich als dieses Unternehmen ausgaben, an das Unternehmen zu melden. Gleichzeitig wurde nach verdächtigen Phishing-Webseiten gesucht. Insgesamt wurden 115.000 E-Mails gemeldet und fast 12.000 Webseiten innerhalb eines knappen Jahres (Februar bis Juli und September bis Dezember 2018) identifiziert. Diese Meldungen waren freiwillig sowie ohne irgendwelche Anreize und die Dunkelziffer der Phishing-Nachrichten im Namen dieses Unternehmens ist sehr wahrscheinlich deutlich größer. Dies verdeutlicht die große Dimension von Phishing-Angriffen aus einer weiteren Perspektive. Die Meldungen erfolgten durch Kundinnen und dadurch wurden E-Mails mehrfach gemeldet. Van der Heijden und Allodi haben dazu die Texte der E-Mails verglichen und nach der Ähnlichkeit zusammengefasst. Insgesamt haben sie damit 1.241 unterschiedliche Texte gefunden [134].

Van der Heijden und Allodi haben den Inhalt untersucht. Sie haben sich dabei auf die Prinzipien zur Be-

einflussung von Menschen nach Cialdini gestützt. Cialdini hat die folgenden Prinzipien entwickelt, um Menschen zu beeinflussen: Reziprozität, Konsistenz, sozialer Beweis, Autorität, Beliebtheit und Knappheit [31]. Diese Prinzipien sollen es, zum Beispiel in der Werbung, ermöglichen, Menschen und deren Handlungen zu beeinflussen [134]. Besonders häufig wurde in den E-Mails das Prinzip der Knappheit, zum Beispiel der Account wird gesperrt, verwendet. Dieses Prinzip ist in den obigen Fallbeispielen häufig erkennbar.

Neben dem Inhalt einer Phishing-Nachricht muss die Herkunft (E-Mail-Adresse oder Webseite) sowie ein Link zur einer URL legitim wirken. Tian et al. haben mehr als 700 Organisationen beobachtet und in einem DNS-Datensatz etwa 600.000 mögliche Angriffsdomains identifiziert. Von diesen waren etwa 300.000 noch aktiv und mehr als 87% leiteten zu der legitimen Webseite der entsprechenden Organisation. Trotzdem konnten sie noch mehr als 850 Phishing-Webseiten identifizieren [132]. Populäre Marken sind sich dieser Angriffe bewusst und registrieren darum aktiv mögliche Phishing-Kandidaten. Es ist aber fraglich, ob dies in kleineren Organisationen ähnlich ist. Insbesondere bei gezielten Angriffen ist dies ein Gefahrenpotential. Simpson et al. haben zu 95% von 270.000 Unternehmen in den Jahren 2009 bis 2019 mindestens eine unregistrierte Domain identifiziert, welche nach ihrem Kriterium ähnlich zu der legitimen Domain ist und potentiell zur Täuschung über die Legitimität einer URL genutzt werden kann [125]. Gleichzeitig wurde bei nur 7% von etwa 250.000 Unternehmen (mit einer *.com*-Domain) mindestens eine registrierte sehr ähnliche Domain, welche potentiell für Angriffe genutzt werden kann, identifiziert. Quinkert et al. haben Domains gesucht, welche zwar ähnlich aussehen, aber Zeichen aus unterschiedlichen Alphabeten benutzen [102]. Angriffe dieser Art werden als homoglyphisch bezeichnet. Innerhalb von einer achtmonatigen Studie konnten sie fast 3.000 homoglyphische Domains für 819 legitime Domains identifizieren. Davon wurden aber nur 28 als Phishing-Angriff kategorisiert. Die Mehrheit der Domains war nur registriert oder geparkt [102].

Die Abgrenzung zu anderen fragwürdigen oder bösartigen Verhalten im Kontext von Domains ist fließend. Ein häufiger Untersuchungsgegenstand sind sogenannte *Typo-squatting* Domains [2, 89, 130, 139]. Diese Domains unterscheiden sich um einen Buchstaben zu einer populären Domain und die Annahme ist häufig, dass diese Domains besucht werden, wenn Nutzerinnen einen Tippfehler bei der Eingabe machen. Das Verhalten dieser Webseiten reicht von nicht erreichbar über Werbung bis hin zu der Abfrage von Zugangsdaten. Unter diesen Webseiten befinden sich somit potentielle Phishing-Webseiten. Agten et al. haben 500 unterschiedliche populäre Domains betrachtet und konnten zu 477 mindestens eine ähnliche und bösartige Domain identifizieren. Dagegen haben nur 156 von den populären Domains ähnliche Domains vorsorglich registriert.

3.5 Zusammenfassung

Phishing-Angriffe sind ein allgegenwärtiges Sicherheitsproblem, wobei die Abgrenzung in der Praxis schwierig ist, weil sich die Angriffstechniken mit anderem fragwürdigen Verhalten im Internet überschneiden.

Auffällig ist, dass Phishing-Angriffe von unterschiedlicher Qualität und teilweise mehrstufig sind. Gezielte Angriffe stellen dabei eine besondere Herausforderung dar.

Das Angriffsziel ist die Täuschung der Herkunft, um eine bestimmte Aktion auszulösen. Im nächsten Kapitel wird der Angriff formalisiert. Anschließend wird die Herkunft einer E-Mail und deren Darstellung genauer betrachtet.

Kapitel 4

Formale Modellierung

Im vorherigen Kapitel wurden Phishing-Angriffe beschrieben und die Gefahren sowie der Kern des Angriffes aus einer Definition abgeleitet. In diesem Kapitel wird aus diesem Gefahrenmodell ein formales Sicherheitsmodell entwickelt und durch mehrere Spiele betrachtet. Das Ziel dieser Formalisierung ist, die Analyse von Darstellungen zu ermöglichen und zu diese zu vergleichen.

4.1 Legitimitätsspiel

Bei einem Phishing-Angriff erhalten Nutzerinnen eine Nachricht und werden aufgefordert, eine sicherheitskritische Aktion zu tätigen, wie zum Beispiel ein Passwort einzugeben. Es wird angenommen, dass die Ausführung solcher sicherheitskritischen Aktionen an eine bestimmte Herkunft der Nachricht gebunden ist. Zum Beispiel wird sehr wahrscheinlich auf den Link einer Nachricht zur Passwortverwaltung nur geklickt, wenn der Ursprung der Nachricht die dazugehörige Webseite ist und dort ein Nutzerkonto vorhanden ist. Es wird angenommen, dass mit der entsprechenden legitimen Partei bereits in der Vergangenheit kommuniziert wurde. Im Beispiel mit dem Nutzerkonto wurde zur Nutzerkontoerstellung bereits per E-Mail kommuniziert. In diesem Fall werden E-Mails ausgetauscht.

Im Folgenden werden die Nutzerin als $\mathcal{H}$, die Angreiferin als $\mathcal{A}$ und eine Darstellung (innerhalb einer Anwendung) als Π, welche aus Gen, R besteht, bezeichnet.

Für die Darstellung wird angenommen, dass es eine Initialisierungsmethode gibt. Diese wird als Gen bezeichnet. Daneben hat die Anwendung noch eine Funktion R zur grafischen Darstellung einer Nachricht m. Die Anwendung kann einen Zustand über vergangene Kommunikation und weiteren Kontext zur E-Mail haben. Dieser Zustand (Σ) ist eine weitere Eingabe für die Darstellungsfunktion R. Dies umfasst die Meta-Informationen, wie das `FROM`-Feld oder den Betreff sowie den Inhalt samt Anhang.

Der Mensch $\mathcal{H}$ wird in den Spielen als eine *Black-Box* simuliert. An $\mathcal{H}$ können unterschiedliche Anfragen

gestellt werden und $\mathcal{H}$ gibt eine Entscheidung auf die Anfrage zurück. In der Praxis sowie in Experimenten ist die menschliche Entscheidung von vielen Parametern und Faktoren abhängig. Eine erste mögliche Entscheidung für einen Menschen im Kontext einer Nachricht ist die Frage nach der Legitimität der dargestellten Nachricht. Diese Entscheidung wird im ersten Sicherheitsspiel betrachtet. Die erste Anfragemöglichkeit an $\mathcal{H}$ wird als legit bezeichnet. Bei dieser Anfrage wird gefragt, ob eine Nachricht legitim ist. Die Antwort kann Ja (1) oder Nein (0) sein.

Ein wichtiger Aspekt dabei ist, dass die Angreiferin eine falsche Identität vortäuscht. Diese vorgetäuschte dritte Partei (falsche Identität) ist der Nutzerin bekannt, wird als legitim eingestuft und somit gibt es eine Kommunikationsvergangenheit, bevor der Angriff stattfindet.

Diese Kommunikationsvergangenheit ist essenzieller Teil des Angriffs und kann formal als bestimmter Zustand vom Menschen aber auch der Anwendung beschrieben werden. Es wird angenommen, dass $\mathcal{H}$ diesen Zustand bei der Beantwortung späterer Fragen berücksichtigt. Weiterhin wird angenommen, dass das Verfahren einen internen Zustand über die Kommunikationsvergangenheit hat. Beispielsweise kennt ein Verfahren die bisherigen Kommunikationspersonen und die ausgetauschten Nachrichten. Im Kontext einer E-Mail-Anwendung ist es naheliegend, denn der Nutzerin wird eine Suchfunktion und der Verlauf einer Konversation (mittels *threading*) zur Verfügung gestellt. Ein Verfahren kann diese Informationen berücksichtigen und der Zustand wird mit Σ bezeichnet. Die Angreiferin kann vor dem eigentlichen Angriff versuchen möglichst viele Informationen zu sammeln. Zusätzliche Informationen für die Angreiferin werden als Leak bezeichnet. Leak ermöglicht eine Unterscheidung zwischen bestimmten Arten des Angriffs, wie zum Beispiel gezielte Angriffe oder eher breitflächige Angriffe, die schwer abzugrenzen sind von Spam. Beispielsweise enthält Leak die E-Mail-Adressen und allgemeine öffentlich zugängliche E-Mails von der legitimen Partei. Eine Standard-E-Mail zum Zurücksetzen eines Passworts ist ein klassisches Beispiel.

Falls kryptographische Funktionen genutzt werden, werden Angriffe gegen diese ausgeschlossen. Weitere Einschränkungen an die Angreiferinnen sind zusätzlich möglich.

Ein Phishing-Angriff ist als Spiel 1 zwischen Angreiferin, Verfahren und Menschen dargestellt. Dazu ist eine Abstraktion zur Simulation der Vergangenheit nötig, denn die Täuschung ist ein wesentlicher Aspekt. Die Funktion Setup stellt das bereit. Im Kontext einer E-Mail-Nutzung wird beispielsweise ein Postfach samt der Kommunikation erstellt und sichergestellt, dass $\mathcal{H}$ den erzeugten Inhalt kennt. Daneben gibt es noch eine weitere legitime Nachricht. Diese dient als Referenz-Nachricht und $\mathcal{H}$ sollte diese Nachricht als legitim einstufen. In diesem Spiel muss $\mathcal{H}$ entscheiden, ob die dargestellte Nachricht legitim ist.

Die Funktion Setup ist sehr mächtig. Im Kapitel 2 wurde gezeigt, dass die E-Mail-Nutzung sehr unterschiedlich ist. Die vergangene Kommunikation von Nutzerinnen kann sich sehr unterscheiden [45, 53, 128]. Nach der formalen Beschreibung von Spiel 1 wird Setup genauer betrachtet. Das Spiel 1 hat als Eingabe die Angreiferin ($\mathcal{A}$), das Verfahren $(\mathrm{Gen}, \mathrm{R})$, $\mathrm{Leak}, \mathrm{Setup}, \mathcal{H}$. Die Ausgabe sind jeweils zwei Bits (b und $\hat{b}$). Dies weicht von der üblichen Notation ab, aber damit können dann vier mögliche Ausgaben vom Spiel einfach dargestellt werden. Diese sind im späteren Verlauf nötig und erweisen sich als

$\underline{\text{Legitim}^{\mathcal{A},\mathcal{H}^{\text{legit}}}_{\text{Setup,Leak,Gen,R}}}$

1 : $\Sigma_{\text{R}}^{0} \leftarrow \text{Gen}()$

2 : $\Sigma_{\text{R}}^{\star}, m_1 \leftarrow \text{Setup}^{\mathcal{H}^{\text{legit}}}(\Sigma_{\text{R}}^{0})$

3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$

4 : $b \leftarrow\!\!\$\ \{0,1\}$

5 : $\hat{b} \leftarrow \mathcal{H}^{\text{legit}}(\text{R}(m_b, \Sigma_{\text{R}}^{\star}))$

6 : **return** $\hat{b}, b$

Sicherheitsspiel 1: In diesem Sicherheitsspiel wird nach der Legitimität einer Nachricht gefragt.

einfache und praktische Formalisierung der Ausgabe auch für weitere Spiele. Andere Varianten sind als Ausgabe sind auch möglich, aber diese war die einfachste und flexibelste.

In dem Spiel 1 gewinnt $\mathcal{A}$ genau dann, wenn gilt $b = 0$ und $\hat{b} = 1$, also $\mathcal{H}$ getäuscht wurde. $\mathcal{H}$ bleibt auf der sicheren Seite genau dann, wenn gilt $b = \hat{b}$. In diesem Sicherheitsspiel werden nicht die genaue Identität und die genaue Aktion berücksichtigt. Die Sicherheitsdefinition ist somit wie folgt:

Definition 5. *[Täuschung über Legitimität von Nachrichten] Eine Darstellung* $\Pi = (\text{Gen}, R)$ *ist* β*-sicher gegen Täuschung einer (durchschnittlichen) Nutzerin* $\mathcal{H}$ *mit* Setup *über legitime Nachrichten, gdw. für alle Angreifer* $\mathcal{A}$ *mit* Leak *gilt:*

$$\Pr[\textit{Legitim}^{\mathcal{A},\mathcal{H}^{\textit{legit}}}_{\text{Setup,Leak,Gen},R} = 1, 0 | b = 0] \leq \beta$$

Die Erfolgswahrscheinlichkeit von $\mathcal{A}$ wird mit β bezeichnet. In dieser Sicherheitsdefinition werden alle Angreiferinnen mit Zugriff auf Leak betrachtet und so die best mögliche Angreiferin. Eine offene Frage ist, ob Verfahren eine Betrachtung aller Angriffe ermöglichen. Im Gegensatz dazu ist es fast unmöglich, Sicherheitsaussagen über alle Nutzerinnen zu treffen. Dies würde erfordern, dass alle Nutzerinnen individuell betrachtet werden. In den Kognitionswissenschaften, der Psychologie und im Kontext der Untersuchung der Benutzbarkeit von Anwendungen werden häufig eher Schätzer, wie der Durchschnittswert, betrachtet [115, 116, 117].

Neben der Definition der Sicherheit ermöglicht das Spiel 1 eine Abschätzung über die Akzeptanz von legitimen Nachrichten. Hierzu wird der Fall betrachtet, dass $\mathcal{H}$ eine legitime Nachricht sieht. Diese Eigenschaft ist ähnlich zu den Anforderungen aus der Kryptographie, dass verschlüsselte Nachrichten mit hoher Wahrscheinlichkeit wieder entschlüsselt werden oder korrekt signierte Nachrichten verifiziert werden können.

Die Akzeptanz einer legitimen Nachricht ist wie folgt definiert:

Definition 6. *[Akzeptanz] Eine Darstellung* $\Pi = (\text{Gen}, R)$ *hat* α*-Akzeptanz, gdw. für eine (durchschnittliche) Nutzerin* $\mathcal{H}$ *mit* Setup *und für alle Angreifer* $\mathcal{A}$ *mit* Leak *gilt:*

$$\Pr[\textit{Legitim}^{\mathcal{A},\mathcal{H}^{\textit{legit}}}_{\text{Setup,Leak,Gen},R} = 1, 1 | b = 1] \geq \alpha$$

Die falsche Zurückweisung einer legitimen Nachricht hat zwar keine direkte Auswirkung auf die Sicherheit, aber ist aus folgenden Gründen problematisch:

1. Beeinträchtigung der Nutzerinnen im Alltag und damit verbunden eine mögliche Einschränkung der Benutzbarkeit der Anwendung.
2. Es besteht die Gefahr, dass bei vielen Fehlern Nutzerinnen die Darstellung nicht mehr beachten.

Neben der Darstellung Π haben $\mathcal{H}$ und Setup einen großen Einfluss auf die Sicherheit der Darstellung, aber beide sind noch sehr abstrakt beschrieben. Die Definitionen 6 und 5 werden in der Definition 7 zusammengefasst.

Definition 7. *[Legitimität von Nachrichten] Eine Darstellung* $\Pi = (\mathrm{Gen}, R)$ *ist* α, β*-sicher gegen Täuschung einer (durchschnittlichen) Nutzerin* $\mathcal{H}$ *mit* Setup *über legitime Nachrichten, gdw. für alle Angreifer* $\mathcal{A}$ *mit* Leak *gilt:*

$$\Pr[\mathit{Legitim}^{\mathcal{A},\mathcal{H}^{legit}}_{\mathrm{Setup},\mathrm{Leak},\mathrm{Gen},R} = 1, 0 | b = 0] \leq \beta$$
$$\Pr[\mathit{Legitim}^{\mathcal{A},\mathcal{H}^{legit}}_{\mathrm{Setup},\mathrm{Leak},\mathrm{Gen},R} = 1, 1 | b = 1] \geq \alpha$$

In der Definition 7 werden die falschen Entscheidungen der Nutzerin als das Maß für die Sicherheit und Benutzbarkeit des Verfahrens betrachtet.

4.1.1 Nutzerin $\mathcal{H}$

Im Gegensatz zu kryptographischen Sicherheitsspielen ist bei der Modellierung von Phishing-Angriffen der Mensch von essentieller Bedeutung. Die Nutzerinnen werden mittels $\mathcal{H}$ simuliert und dazu wird ein *Black-Box*-System angenommen, welches Antworten auf bestimmte Anfragen gibt. Im Spiel 1 war die Anfrage an $\mathcal{H}$, ob eine Nachricht legitim ist. Die Anfrage wurde mit legit bezeichnet. In Anlehnung an die kryptographische Sicherheitsspiele kann $\mathcal{H}$ damit als ein Orakel betrachtet werden.

Die Antworten von $\mathcal{H}$ müssen dabei nicht korrekt sein. Zwar trifft $\mathcal{H}$ die Entscheidung, eine Nachricht als legitim einzustufen, aber es können fehlerhafte oder falsche Entscheidungen sein.

$\mathcal{H}$ kann betrachtet werden als das zufällige Auswählen einer Nutzerin aus der Gesamtpopulation der Nutzerinnen. In diesem Fall wird die Wahrscheinlichkeit über den Spielausgang durch die wiederholte Anwendung der Experimente bestimmt. Eine andere Sichtweise kann sein, dass aus den unterschiedlichen Nutzerinnen eine modellhaften Repräsentantin gebildet wird und quasi als Standardnutzerin fungiert. Hierbei wird bereits deutlich, dass $\mathcal{H}$ aus unterschiedlichen Perspektiven betrachtet werden kann. Die Simulation von $\mathcal{H}$ kann durch empirische Daten simuliert oder durch (kognitive) Modelle erfolgen. Dies

wird zunächst offen gelassen, aber kann bei der Untersuchung eines konkreten Verfahrens berücksichtigt werden.

Eine wichtige Eigenschaft von $\mathcal{H}$ ist, dass zwei gleiche Anfragen zu unterschiedlichen Ergebnissen führen können. Durch diese Eigenschaft wird die fehleranfällige Entscheidung noch weiter erhöht und teilweise schwer nachvollziehbar. Das Ziel des Verfahrens ist, möglichst viele sichere Entscheidungen herbeizuführen.

Bei einem Phishing-Angriff erfolgt eine Identitätstäuschung und diese Identitäten sind $\mathcal{H}$ bekannt. Aus dieser Betrachtungsweise hat $\mathcal{H}$ einen inneren Zustand und entscheidet nach diesem, ob eine Nachricht legitim ist. Wenn $\mathcal{H}$ die angegebene Herkunft einer Nachricht, zum Beispiel Google, nicht kennt, dann ist jegliche Frage nach der Legitimität dieser Nachricht sinnlos. Dieser innere Zustand muss in diesem Spiel sichergestellt werden. Für Phishing-Sicherheitsspiele wird deshalb angenommen, dass bereits eine vorherige Kommunikation erfolgte.

In der Realität ist die Frage, ob eine Nachricht legitim ist, doch eine sehr mächtige Anfrage an einen Menschen und wird sehr unterschiedlich beantwortet. Die Entscheidung wird von vielen unterschiedlichen Faktoren beeinflusst, wie zum Beispiel:

1. Vertrautheit der Anwendung
2. Kenntnisse und Fachwissen über die Anwendung, Kommunikationsprotokoll und IT im Allgemeinen
3. Sensibilisierung gegenüber Angriffen und insbesondere Phishing-Angriffen
4. Erfahrungen von vergangenen Angriffen
5. Nutzung der Anwendung und Art der Kommunikation
6. Kontext der Kommunikation
7. Tagesform und aktuelle äußere Einflüsse
8. Demographische Faktoren

In dieser Konsequenz ist die Modellierung der Anfrage von legit in diesem Spiel eine sehr mächtige und komplexe Anfrage, welche wahrscheinlich großen Schwankungen ausgesetzt ist. Aus diesem Grund werden Sicherheitsspiele mit weniger komplexen Anfragen betrachtet. Damit können Sicherheitsmechanismen grundsätzlicher untersucht werden.

4.1.2 Erzeugung vom Kontext

Phishing-Angriffe erfolgen in einem gewissen Kontext mit einer Vergangenheit und dieser Kontext wird mittels Setup in dem Sicherheitsspiel simuliert. Dieser Kontext wird einerseits durch die Nutzerin $\mathcal{H}$ als auch durch die vergangene Kommunikation beeinflusst. Diese vorherige Kommunikation erfolgte in der

Regel über die Anwendung. Die Anwendung kann die vorherige Kommunikation speichern, verarbeiten und für die Sicherheitsanalyse zukünftiger Kommunikation nutzen. Es ist denkbar, dass je nach Kontext unterschiedliche Mechanismen unterschiedliche Auswirkungen auf die Sicherheit haben. Der Kontext wird im Wesentlichen durch die vergangene Kommunikation bis zu einem Angriff bestimmt. Hierbei wird deutlich, dass es viele unterschiedliche Varianten gibt, und einige Faktoren dabei sind:

1. Anzahl der Kommunikationspartnerinnen, deren (soziale) Beziehungen und Bedeutung
2. Anzahl der ausgetauschten Nachrichten
3. Art und Häufigkeit des Nachrichtenaustausches wie zum Beispiel nur Empfang von Nachrichten, Dialoge, Diskussionen mit unterschiedlichen Kommunikationspartnerinnen
4. Häufigkeit und Art von ausgesetzten Angriffen

Diese Liste ist unvollständig und zeigt die Komplexität vom Kontext und es ist vorstellbar, dass vieles davon Auswirkungen auf die Entscheidung von $\mathcal{H}$ hat. Eine besondere Bedeutung haben die Dienstleister zur Bereitstellung des Kommunikationsnetzwerkes. Diese können bereits Nachrichten filtern und aussortieren, welche sowohl die Anwendung als auch Nutzerinnen nicht erhalten oder dargestellt bekommen. Ein wichtiger Aspekt ist dabei insbesondere die Häufigkeit von für Nutzerinnen dargestellten Angriffen.

Mit Setup wird im Sicherheitsspiel die Ausgangssituation mit legitimen Nachrichten für den Angriff erzeugt. Dies beinhaltet explizit, dass der Mensch $\mathcal{H}$ die Darstellung Π genutzt hat und bestimmte Darstellungen gesehen hat. Aus diesem Grund bekommt Setup Zugriff auf $\mathcal{H}$. Dies ist notwendig, weil $\mathcal{H}$ im Sicherheitsspiel die endgültige Entscheidung trifft und mit der Entscheidung von $\mathcal{H}$ entschieden wird, ob $\mathcal{A}$ im Sicherheitsspiel gewinnt oder nicht. Diese Entscheidung passiert auf Grundlage der vergangenen Kommunikation und diese muss im Sicherheitsspiel berücksichtigt werden. Setup gibt eine Nachricht vom Ziel des Angriffs zurück, welche als Grundlage für einen Angriff genutzt werden kann, und den Zustand vom Darstellungsverfahren, nachdem $\mathcal{H}$ bereits Π genutzt hat.

Gleiche Vergangenheit

Die Aufgabe von Setup ist die Erzeugung einer gemeinsamen Vergangenheit von $\mathcal{H}$ und Π auf deren Grundlage die Angreiferin $\mathcal{A}$ einen Angriff durchführen kann.

Eine einfache konkrete Ausprägung von Setup kann wie folgt aussehen:

1. Erzeuge n unterschiedliche Herkünfte.
2. Wähle k Herkünfte, welche wiederholt dargestellt werden, aus.
3. Erzeuge zu jedem Kontakt eine Nachricht und zeige diese $\mathcal{H}$ mittels Π.
4. Falls eine Herkunft wiederholt dargestellt werden soll, dann zeige nach mehr als sieben Nachrichten als Herkunft einer weiteren Nachricht an.[1]

[1] Damit wird angenommen, dass die Nutzerin sich die Nachrichten oder Teile davon nicht im Kurzzeitgedächtnis merken kann.

5. Wähle das Ziel t als eine der ersten l Herkünfte zufällig.

6. m_1 ist eine neue Nachricht vom Ziel t.

Die obige Ausprägung von Setup ist relativ einfach und überschaubar.[2] $\mathcal{H}$ sieht n unterschiedliche Herkünfte einer Nachricht, aber diese wiederholen sich nicht oft. Unter der Annahme, dass $\mathcal{H}$ effizient Nachrichten bearbeiten möchte, wird jeder Darstellung nicht sehr lange Aufmerksamkeit gewidmet und damit gibt es nur wenige Möglichkeiten von $\mathcal{H}$, mit der Darstellung einer Herkunft und deren Nachrichten vertraut zu werden. In einem realen Nachrichtenverlauf einer Person wiederholen sich Nachrichten einiger Herkünfte häufiger. In diesem Fall hat $\mathcal{H}$ mehr Möglichkeiten, mit der Darstellung einer Herkunft und deren Nachrichten vertraut zu werden. Die obige Ausprägung von Setup ist damit eine besondere Herausforderung für $\mathcal{H}$. In einem konkreten Experiment muss die Unterteilung zwischen Setup und Angriff nicht so deutlich sein, denn in einem Nachrichtenstrom ist der Übergang fließend.

Die Informationsbox 4.1.2 ordnet ein praktisches Beispiel aus dem vorherigen Kapitel dem formalen Sicherheitsspiel zu und ordnet es bisherigen wissenschaftliche Studien zu.

[2] Diese Ausprägung wird in Kapitel 4 aufgegriffen und aus den kognitiven Experimenten abgeleitet.

Legitimitätsspiel(4.1.2)

Praktisches Beispiel

Ein E-Mail-Provider erhält von einem anderen E-Mail-Provider eine E-Mail und soll entscheiden, ob diese E-Mail legitim ist oder nicht. Bei einer (voraussichtlich) legitimen E-Mail wird diese den Nutzerinnen zur Verfügung gestellt. Nutzerinnen stellen sich die Frage nicht am Anfang, sondern, wie Billy Rinehart, wenn sie bei der Bearbeitung von E-Mails skeptisch werden.

Wissenschaftliches Beispiel

Zur Evaluation von ihrem Trainingsspiel zur Erkennung von Phishing zeigten Wen et al. ihre Studienteilnehmerinnen bei 11 Phishing-Angriffen und 9 legitimen E-Mails, ob die dargestellte E-Mail legitim oder eine Phishing-E-Mail ist [143].

Eignung

Vorteile	**Nachteile**
Einfache Frage	Keine alltägliche erste Frage für Nutzerinnen. Für Studienteilnehmerinnen ist das Ziel der Studie offensichtlich.

Diese Fragestellung ist zur Untersuchung von Schulungsmaterial in einer Labor- oder Online-Studie geeignet, weil der Untersuchungsgegenstand offensichtlich ist. Für die Untersuchung von einem Verfahren in einer Anwendung stellt es keine Alltagssituation dar.

4.1.3 Sicherheitsspiel als Experiment

Eine andere Perspektive auf das Spiel 1 ist die Betrachtung als eine formale und allgemeine Beschreibung von Experimenten. Das Experiment unterteilt sich dabei in eine Lernphase und Testphase. Die Lernphase ist im Sicherheitsspiel Setup und die Testphase ist die Präsentation der Nachricht m_b.

Die Teilnehmerinnen in einer Studie sind die Konkretisierung von $\mathcal{H}$ und diese bewerten die dargestellten Nachrichten als legitim oder nicht.

In diesen Experimenten wird Setup in der Regel von der Versuchsleitung erzeugt und simuliert ein typisches Postfach. Gleichzeitig wählt die Versuchsleitung einen oder mehrere Angriffe aus. Bei Betrach-

b	$\hat{b}$	Auswertung im Spiel 1	Ergebnis
1	1	$\mathcal{H}^{\text{legit}}(\mathrm{R}(m_1), \Sigma_{\mathrm{R}}^{\star}) = 1$	korrekte Akzeptanz
1	0	$\mathcal{H}^{\text{legit}}(\mathrm{R}(m_1), \Sigma_{\mathrm{R}}^{\star}) = 0$	falsche Zurückweisung
0	1	$\mathcal{H}^{\text{legit}}(\mathrm{R}(m_0), \Sigma_{\mathrm{R}}^{\star}) = 1$	falsche Akzeptanz
0	0	$\mathcal{H}^{\text{legit}}(\mathrm{R}(m_0), \Sigma_{\mathrm{R}}^{\star}) = 0$	korrekte Zurückweisung

Tabelle 4.1: Die Fehler sind eine falsche Akzeptanz oder Zurückweisung, wobei eine falsche Akzeptanz eine besondere Gefahr für Nutzerinnen darstellt.

tung vom Sicherheitsspiel wird deutlich, dass es eine große Herausforderung ist, den besten Angriff auszuwählen und durchzuführen.

Eine weitere Herausforderung ist, die gemeinsame Vergangenheit zwischen dem Verfahren und den Teilnehmerinnen mittels Setup zu erzeugen. Dies bedeutet insbesondere, dass die Teilnehmerinnen mit den Darstellungen vertraut sein. Eine andere Alternative ist, dass ein Experiment als Feldstudie durchgeführt wird, dann ist Setup nicht von der Versuchsleitung bestimmt und schwer kontrollierbar. In allen Fällen beeinflusst Setup die Messungen im Experiment und hat somit Auswirkungen auf die Sicherheitsanalyse.

Für den Ausgang des Spieles 1 mit der Ausgabe von $\hat{b}$ und b gibt es vier verschiedene Ausgänge, welche in der Tabelle 4.1.3 dargestellt werden.

Hierbei wird deutlich, dass es ein Experiment über die Entdeckung von verdächtigen Anzeichen (oder abstrakt von Signalen im Sinne der Signalentdeckungstheorie [146]) ist und vier Ausgänge hat. Ein Treffer ist die korrekte Einstufung einer legitimen Nachricht und die Zurückweisung einer nicht legitimen Nachricht von $\mathcal{A}$ ist eine korrekte Zurückweisung. Die falsche Zurückweisung einer legitimen Nachricht ist damit ein Akzeptanzproblem und somit eher ein Gebrauchstauglichkeitsproblem. Eine legitime Nachricht wird dann nicht bearbeitet und damit verhindert das Verfahren die Kommunikation und damit den Gebrauch der Anwendung. Die falsche Akzeptanz einer nicht legitimen Nachricht ist ein erfolgreicher Angriff und somit ein Sicherheitsproblem.

Durch die vier möglichen Ausgänge (korrekte Akzeptanz, falsche Zurückweisung, falsche Akzeptanz, korrekte Zurückweisung) sind allgemeine statistische Interpretationen und Betrachtungsweisen, welche aus klinischen Studien bekannt sind [22], möglich. Wichtige Kennzahlen sind dabei die korrekte Akzeptanzrate (*true positive rate, TPR*) und die falsche Akzeptanzrate (*falsche positive rate, FPR*). Sei tp die Anzahl der korrekten Akzeptanzen (*true positives*), fr die Anzahl der falschen Zurückweisungen (*false rejections*), fp die Anzahl der falschen Akzeptanzen (*false positives*), tr die Anzahl der korrekten Zurückweisungen (*true rejections*). Dann sind die TPR, FPR wie folgt definiert:

$$TPR = \frac{tp}{tp + fr} = \Pr[\mathcal{H}^{\text{legit}}(\mathrm{R}(m_b), \Sigma_{\mathrm{R}}^{\star}) = 1 \mid b = 1] = \alpha$$
$$FPR = \frac{fp}{fp + tr} = \Pr[\mathcal{H}^{\text{legit}}(\mathrm{R}(m_b), \Sigma_{\mathrm{R}}^{\star}) = 1 \mid b = 0] = \beta$$

Die obigen Gleichungen stellen ebenso die theoretische Beziehung zwischen den empirischen Raten und den Wahrscheinlichkeiten aus dem Sicherheitsspiel dar. Wenn die Fehlerrate experimentell bestimmt wird, dann ist dies eher eine Schätzung für die Wahrscheinlichkeit im Sicherheitsspiel.

Weitere Kennzahlen sind unter anderem die Sensitivität oder Spezifität und diese ermöglichen den Vergleich zwischen Verfahren. In dem konkreten Anwendungsfall ist es insbesondere interessant, ob ein Verfahren **brauchbar** im Sinne von einer guten Unterstützung der Entscheidung des Menschen ist. Aus der Untersuchung von medizinischen diagnostischen Tests ist die *likelihood ratio positive (LHR+)*, welche wie folgt definiert ist [41], vertraut:

$$LHR+ = \frac{TPR}{FPR} = \frac{\alpha}{\beta}$$

Wenn $LHR+ = 1$ ist, bedeutet dies, dass das Verfahren keine Trennung zwischen legitimen und bösartigen Nachrichten ermöglicht. Bei der Beurteilung von diagnostischen Test wird dieses verwendet [41]. Dujardin et al. diskutieren die Vorteile sowie die Nachteile von unterschiedlichen Kennzahlen [41].

Brauchbarkeit eines Verfahrens

Ein Verfahren ist nicht brauchbar, wenn die $LHR+ \leq 1$ bzw. $\alpha \leq \beta$ ist. Sei $\epsilon = \frac{\alpha}{\beta}$. ϵ wird als Brauchbarkeit bezeichnet.

Bei der Betrachtung des Spiels als Experiment wird die Sicherheit empirisch geschätzt. Hierfür muss das Experiment mehrfach durchgeführt werden und die grundsätzliche Frage ist, wie häufig die Studienteilnehmenden die Legitimität einer Nachricht bewerten.

Zur Reduzierung der nötigen Anzahl an Studienteilnehmerinnen kann eine Person mehrfach verschiedene Nachrichten bewerten und der Setup-Teil muss nur einmal durchgeführt werden. Die Häufigkeit eines Angriffs beeinflusst aber die Aufmerksamkeit einer Person und damit das Ergebnis eines Experiments. Dies ist ein gängiges Problem bei Experimenten, die auf der Signal-Entdeckungs-Theorie [146] aufbauen. In diesem Experiment kommt allerdings erschwerend hinzu, dass im Kontext von E-Mails im Vergleich zu den eingehenden E-Mails der Eingang einer bösartigen E-Mail ein seltenes Ereignis ist.

Näher an dem formalen Sicherheitsspiel hingegen ist es, wenn jede Person nur eine Nachricht bewertet. Die Anzahl der teilnehmenden Personen steigt in diesen Experimenten drastisch, aber die Aufmerksamkeit der Person wird durch die Häufigkeit von Angriffen nicht beeinflusst. Bei einer tatsächlichen Experimentdurchführung ist dies zu berücksichtigen.

Daneben ist explizite Legitmitätsbewertung einer Nachricht eine nicht alltägliche Aufgabe. Im Folgenden werden weitere Sicherheitsspiele betrachtet, welche näher an der normalen Bearbeitung einer E-Mail sind.

4.2 Spiel mit Herkunft und Aktion

Das Ziel des nächsten Spieles (Spiel 2) ist die Definition unter der Einbeziehung der legitimen dritten Partei und der getätigten Aktion. Für das nächste Spiel wird darum eine neue Anfrage an $\mathcal{H}$ benötigt. Die Anfrage nach der Herkunft wird mit origin bezeichnet und die Anfrage nach der sicherheitskritischen Aktion mit `sudo`. In einer legitimen Nachricht passt die sicherheitskritische Aktion zu der Herkunft der Aufforderung zur Aktion. Für eine legitime, sicherheitskritische Aktion, wie zum Beispiel der Besuch einer bestimmten Webseite zur Passworteingabe, muss die Herkunft der Nachricht legitim sein. Die legitime Herkunft einer Nachricht wird mit φ und die legitime, sicherheitskritische Aktion, wie zum Beispiel einen Link zu öffnen oder ein Passwort einzugeben, mit γ bezeichnet. Die Angreiferin $\mathcal{A}$ hat das Ziel, in beiden die Nutzerin $\mathcal{H}$ zu täuschen. Die Überprüfung der Gleichheit zwischen zwei Aktionen bzw. Herkünften ist 1 (falls diese gleich sind) und ansonsten 0.

Das Spiel 2 ist ähnlich im Ablauf zum Spiel 1 und wie folgt definiert:

$\underline{\text{Phishing}^{\mathcal{A},\mathcal{H}^{\text{origin \& sudo}}}_{\text{Setup,Leak,Gen,R}}}$

1 : $\Sigma_{\text{R}}^{0} \leftarrow \text{Gen}()$
2 : $\Sigma_{\text{R}}^{\star}, m_1, \varphi, \gamma \leftarrow \text{Setup}^{\mathcal{H}^{\text{origin \& sudo}}}(\Sigma_{\text{R}}^{0})$
3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{\varphi}, \hat{\gamma} \leftarrow \mathcal{H}^{\text{origin \& sudo}}(\text{R}(m_b, \Sigma_{\text{R}}^{\star}))$
6 : $\hat{b} \leftarrow \varphi = \hat{\varphi} \wedge \gamma = \hat{\gamma}$
7 : **return** $\hat{b}, b$

Sicherheitsspiel 2: In diesem Sicherheitsspiel wird die abgeleitete Herkunft und die getätigte Aktion untersucht.

Aus der Definition von Phishing geht hervor, dass die Angreiferin ihr Ziel nur durch die Täuschung über die Herkunft erreicht. Für einen erfolgreichen Angriff ist darum ein notwendiges Kriterium, dass sowohl die Aktion durchgeführt als auch über die Identität getäuscht wird. Dies wird durch die Konjunktion in Zeile 6 vom Spiel hervorgehoben.

Eine implizite Annahme über Spiel 2 ist, dass m_1 eine legitime Nachricht von φ ist und zu Aktion γ auffordert. Die Aktion γ ist eine Aktion, welche auf Weisung von φ durchgeführt wird. Daraus ergibt sich folgende Sicherheitsdefinition:

Definition 8. *[Schutz gegen Phishing] Eine Darstellung* $\Pi = (\text{Gen}, R)$ *ist* α, β*-sicher gegen Täuschung einer (durchschnittlichen) Nutzerin* $\mathcal{H}$ *mit* Setup *über Herkunft und Aktion einer Nachricht, gdw. für alle Angreifer* $\mathcal{A}$ *mit* Leak *gilt:*

$$\Pr[\textit{Phishing}^{\mathcal{A},\mathcal{H}^{\textit{origin \& sudo}}}_{\text{Setup,Leak,Gen},R} = 1, 0 | b = 0] \leq \beta$$
$$\Pr[\textit{Phishing}^{\mathcal{A},\mathcal{H}^{\textit{origin \& sudo}}}_{\text{Setup,Leak,Gen},R} = 1, 1 | b = 1] \geq \alpha$$

In dieser Definition ist β die Erfolgswahrscheinlichkeit für einen Angriff und $1 - \alpha$ ist die Wahrscheinlichkeit einer falschen Zurückweisung der korrekten Nachricht. Die Wahrscheinlichkeit α impliziert die Plausibilität von der geforderten Aktion γ durch die legitime Herkunft φ. Wenn $\mathcal{H}$ bereits die Aktion γ nach Aufforderung von φ nicht durchführt, dann ist die Wahl von γ und φ nicht geeignet, um die Sicherheit zu untersuchen. Denn eine geringe Angriffswahrscheinlichkeit hat nur eine geringe Aussagekraft über die Sicherheit vom Verfahren, wenn die geforderte Aktion selbst bei der legitimen Partei nicht durchgeführt wird. Beispielsweise ist es nicht plausibel, dass Nutzerinnen ihr PayPal-Nutzerkonto löschen, wenn PayPal sie dazu auffordert. In einem Angriff mit der legitimen Partei PayPal ist diese Aktion (Löschung vom Nutzerkonto) damit ebenso ungeeignet. Die Aktion und die Herkunft müssen zusammenpassen.

Zusammenspiel zwischen Herkunft und Aktion

Zur Untersuchung eines erfolgreichen Angriffs mit einer hohen Erfolgswahrscheinlichkeit nach dem Spiel Phishing ist nötig, dass die Aufforderung der legitime Herkunft φ zur Aktion γ plausibel ist.

Die Tabelle 4.2 stellt die möglichen Ausgänge des Sicherheitsspiels 2 dar. Das Ereignis einer falschen Zurückweisung (*false rejection, FR*) bedeutet, dass Nutzerinnen die dargestellte Nachricht weder vom Inhalt mit der gewünschten Aktion noch von der Herkunft mit der gewünschten Herkunft verbinden. Die Ursache kann sein, dass entweder gewählte Nachricht m nicht zu γ und φ passt, die Darstellung dies nicht ermöglicht oder der Kontext der Nutzerinnen die Assoziation verhindert. Falls die Darstellung der Nachricht das größte Hindernis bei der falschen Zurückweisung ist, dann deutet dies auf einen negativen Einfluss auf die Benutzbarkeit der Anwendung hin. Die Anwendung erfüllt nicht (immer) ihren Zweck und hindert eventuell Nutzerinnen an der sinnvollen Bearbeitung der Nachricht.

Die Ereignisse einer korrekten Akzeptanz (*true hit, TH*) und einer korrekten Zurückweisung (*true rejection, TR*) sind wünschenswerte Ausgänge in dem Experiment. Im ersten Fall bedeutet es, dass Nutzerinnen eine Nachricht der richtigen Herkunft zugeordnet haben und die entsprechende Aktion getätigt haben. Bei einer korrekten Zurückweisung wurde die Angriffsnachricht nicht der Herkunft zugeordnet und die Aktion wurde nicht getätigt. Die Nutzerinnen wurden damit nicht erfolgreich angegriffen und waren nicht verwundbar gegen den Angriff. Beide Ereignisse sind positiv in Bezug auf die Sicherheit und Benutzbarkeit der Anwendung.

Das Ereignis PH bedeutet, dass zwar die Herkunft der Nachricht korrekt erkannt wurde, aber die Aktion nicht getätigt wurde. Dieses Ereignis ist damit eher ungewöhnlich und es gibt mehrere mögliche Erklärungsansätze. Eine Möglichkeit ist, dass die Konstellation aus Nachricht m, γ und φ nicht zusammen passt. Das bedeutet, dass entweder eine Nachricht von φ nicht die Aktion γ auslöst oder die Nachricht m eben nicht die Aktion γ auslöst. Dies kann abhängig von den Nutzerinnen und dem jeweiligen Kontext sein. Eine Interpretation ist, dass der Versuchsaufbau fehlerhaft ist. Eine andere Alternative ist, dass die Darstellung für die Aktion hindert, die Aktion auszuführen. Dann kann dies ein Hindernis der Benutzbar-

b	$\varphi = \hat{\varphi}$	$\gamma = \hat{\gamma}$	$\mathcal{A}$ gewinnt	Ereignis
1	0	0		falsche Zurückweisung (*FR*)
1	1	0		partielle korrekte Akzeptanz (*PH*)
1	0	1		unentschlossene Nutzerinnen (*U*)
1	1	1		(korrekte Akzeptanz *TH*)
0	0	0	0	(korrekte Zurückweisung *TR*)
0	1	0	?	partielle Zurückweisung (*PR*)
0	0	1	?	unentschlossene Nutzerinnen (*U*)
0	1	1	1	falsche Akzeptanz (*FH*)

Tabelle 4.2: Für das Sicherheitsspiel 2 gibt es verschiedene Ausgänge. Durch die zwei verschiedenen Beobachtungen gibt es uneindeutige Ergebnisse.

keit sein.

Das Ereignis PR bedeutet, dass der Angriff partiell zwar erfolgreich war, aber doch nicht endgültig. Die Nutzerinnen wurden über die Herkunft der Nachricht getäuscht, aber haben nicht die Aktion ausgeführt. In diesem Fall sind Nutzerinnen zwar noch sicher, aber waren partiell eben angreifbar. Einige Ursachen für die Nicht-Ausführung der Aktion sind analog zu dem Ereignis PH zu betrachten. Des Weiteren kann es sein, dass bei einer leichten Variation des Inhalt der Nachricht der Angriff doch noch erfolgreich sein kann.

Das Ereignis U bedeutet, dass zwar die Aktion getätigt wurde, aber die dazu passende Herkunft nicht erkannt wurde. Die Aktion wurde ohne die richtige Herkunft getätigt und damit ist die Herkunft kein notwendiges Kriterium für die Aktion. Die Aktion war damit nicht sensitiv und abhängig von der Herkunft der Nachricht. Damit ist entweder die Aktion nicht sicherheitskritisch oder die Einstufung durch die Nutzerinnen ist fraglich. Die Interpretation der Sicherheit der Darstellung ist damit nur schwierig möglich und der Experimentalaufbau ist zu hinterfragen. Dies als einen erfolgreichen Angriff zu werten, ist damit fraglich.

Die partiellen und uneindeutigen Ereignisse PH (*partial hit*), PR (*partial rejection*) und U (*undecided*) ermöglichen einen großen Spielraum für mögliche Interpretationen. Es ist besonders kritisch, dass nicht eindeutig klar ist, ob der Versuchsaufbau, Kontext oder das Sicherheitsverfahren wesentlichen Einfluss hat. Ein Ausweg hieraus ist die getrennte Untersuchung der unterschiedlichen Aspekte eines Angriffs. Dies entspricht der klassischen grafischen Unterteilung einer Anwendung. Häufig wird hierbei zwischen einem Bereich über die Herkunft einer Nachricht, zum Beispiel das `From`-Feld, und dem eigentlichen Inhalt der Nachricht unterschieden. Bei künftigen Sicherheitsspielen wird dies berücksichtigt und der Fokus ist die Darstellung der Herkunft einer Nachricht.

Phishingspiel

Praktisches Beispiel

Billy Rinehart aus dem Beispiel aus dem vorherigen Kapitel hat sich folgende Fragen gestellt:

1. Ist die E-Mail von Google?
2. Soll der Anweisung zur Passwortänderung gefolgt werden?

Diese Fragen hat er direkt an seine IT-Abteilung gestellt, weil er sich anscheinend unsicher war. Sie passte in den Kontext, weil er tatsächlich ein Google-Konto hat. Es ist plausibel, dass er das Passwort nur im Glauben einer Anweisung von Google geändert hätte und nicht durch eine andere beliebige Organisation.

Wissenschaftliches Beispiel

In Feldstudien mit Organisationen wird ein Phishing-Angriff mittels einer E-Mail simuliert und die Klicks auf den Link gemessen [52, 79]. Eine Annahme ist wahrscheinlich, dass ein Klick auf den Link erfolgt, wenn Nutzerinnen die Herkunft der E-Mail als legitim einstufen.

Unterschied zum vorherigen Spiel

Im Gegensatz zum Sicherheitsspiel 1 (Legitimitätsspiel) ist die Aufgabe an Nutzerinnen sehr nahe an der alltägliche Bearbeitung einer Nachricht. In einem Experiment muss der Bezug zum Phishing darum nicht sofort offensichtlich sein.

Eignung

Vorteile	**Nachteile**
Alltag der Nutzerinnen	Phishing-Angriff muss auf die Nutzerinnen angepasst sein
Aufgabe ist Teil der Bearbeitung einer Nachricht	Anwendung muss im Alltag genutzt werden
	Viele Faktoren können das Ergebnis beeinflussen

Die Feldstudie untersuchen nicht nur ein Verfahren bzw. eine Anwendung, sondern die Anwendung in Kombination mit einer spezifischen Gruppe von Nutzerinnen. Es ist zur Untersuchung eines gesamten Systems einer Organisation mit deren genutzter Anwendung und Nutzerinnen als empirischen Experiment geeignet.

4.3 Herkunft der Nachricht

Die Täuschung über die Herkunft einer Phishing-Nachricht ist ein wesentlicher Teil eines erfolgreichen Angriffes. Das letzte Spiel (Spiel 2) wird modifiziert und auf die Täuschung über die Herkunft einer Nachricht eingeschränkt. Damit ergibt sich das Sicherheitsspiel 3.

$\underline{\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}}}$

1 : $\Sigma_{\text{R}}^{0} \leftarrow \text{Gen}()$
2 : $\Sigma_{\text{R}}^{\star}, m_1, \varphi \leftarrow \text{Setup}^{\mathcal{H}^{\text{origin}}}(\Sigma_{\text{R}}^{0})$
3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{\varphi} \leftarrow \mathcal{H}^{\text{origin}}(\text{R}(m_b, \Sigma_{\text{R}}^{\star}))$
6 : **return** $\varphi = \hat{\varphi}, b$

Sicherheitsspiel 3: In diesem Sicherheitsspiel wird nur die von $\mathcal{H}$ abgeleitete Herkunft betrachtet.

Ausgehend von Spiel 3 ist die Sicherheit gegen Täuschungen über die Herkunft einer Nachricht, wie folgt definiert:

Definition 9. *[Schutz gegen Herkunftstäuschung] Eine Darstellung* $\Pi = (\text{Gen}, R)$ *ist* α, β*-sicher gegen Täuschung einer (durchschnittlichen) Nutzerin* $\mathcal{H}$ *mit* Setup *über Herkunft und Aktion einer Nachricht, gdw. für alle Angreifer* $\mathcal{A}$ *mit* Leak *gilt:*

$$\Pr[\textit{Origin}^{\mathcal{A},\mathcal{H}^{\textit{origin}}}_{\text{Setup,Leak,Gen,}R} = 1, 0 | b = 0] \leq \beta$$
$$\Pr[\textit{Origin}^{\mathcal{A},\mathcal{H}^{\textit{origin}}}_{\text{Setup,Leak,Gen,}R} = 1, 1 | b = 1] \geq \alpha$$

Durch Nicht-Berücksichtigung der Aktion und bei Fixierung der restlichen Parameter ergibt sich folgende Abschätzung zwischen den beiden Sicherheitsspielen 2 und 3:

$$\Pr[\text{Phishing}^{\mathcal{A},\mathcal{H}^{\text{origin \& sudo}}}_{\text{Setup,Leak,Gen,R}} = 1, 0 | b = 0] \leq \Pr[\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}} = 1, 0 | b = 0]$$

Damit bietet Origin eine allgemeine Möglichkeit als eine obere Schranke bei der Abschätzung der Erfolgswahrscheinlichkeit eines Angriffs. Die Sicherheit einer Anwendung kann mit einem Schwerpunkt auf die Erkennung der Herkunft abgeschätzt werden.

Herkunftsspiel

Praktisches Beispiel

Billy Rinehart aus dem vorherigen Kapitel hat sich gefragt, ob die E-Mail von Google ist. Im Alltag wird diese Frage immer gestellt, um eine E-Mail in den Kontext zu setzen.

Wissenschaftliches Beispiel

Ein wissenschaftliches Experiment in einer Anwendung nach der Herkunft einer E-Mail gefragt wurde, konnte nicht gefunden werden. Dagegen gibt es Studien, in denen Nutzerinnen eine Anwendung erhalten und bei einer Warnung über eine URL, also die Herkunft eines Links oder einer Webseite, erhalten und im Anschluss wird gemessen, ob die Nutzerinnen die Warnung berücksichtigt haben [43, 99]. Petelka et al. haben gemessen, ob ein Link trotz Warnung angeklickt wurde [99]. Damit wurde indirekt die Herkunft der URL gemessen.

Unterschied zum vorherigen Spiel

Im Gegensatz zum Sicherheitsspiel 2 (Phishingspiel) wird die Bedingung von Aktion und Herkunft aufgelöst. Der Kontext von dem Angriff wird damit vereinfacht.

Eignung

Vorteile	**Nachteile**
Einfacher experimenteller Aufbau (zum Beispiel Labor-, Onlinestudien)	Die Darstellung von anderen Aspekten der Nachricht beeinflussen das Ergebnis

Dieses Sicherheitsspiel ist geeignet, um die Bestimmung der Herkunft in einer Anwendung zu untersuchen. Während oder nach der Entwicklung einer Anwendung ist es eine wichtige Untersuchung. Sie kann darum insbesondere von entsprechenden Softwareherstellern durchgeführt werden.

4.4 Einschränkung auf die Herkunftsdarstellung

Zur Darstellung der Herkunft einer Nachricht wird oftmals ein bestimmter Bereich der Darstellung verwendet. Dieser Bereich der Darstellung wird mit R_S bezeichnet. Im Kontext einer E-Mail-Anwendung ist

dies zum Beispiel die angezeigte `From`-Adresse. Diese kann um die Darstellung einer digitalen Signatur ergänzt werden.

Im obigen Spiel wird statt R nur R_S verwendet. Es wird angenommen, dass R_S dazu die gesamte Nachricht enthält. Dies ist aus formalistischen Gründen nötig, weil $\mathcal{A}$ wieder eine Nachricht ausgeben muss, aber aus praktischen Gründen ist dies sinnvoll. Beispielsweise ist zur Darstellung die Verifikation einer Nachricht nötig.

Auf Basis dieser Überlegungen ergibt sich folgendes Spiel 4:

$\underline{\mathrm{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\mathrm{Setup},\mathrm{Leak},\mathrm{Gen},\mathrm{R}_S}}$

1 : $\Sigma^0_{\mathrm{R}_S} \leftarrow \mathrm{Gen}()$

2 : $\Sigma^\star_{\mathrm{R}_S}, m_1, \varphi \leftarrow \mathrm{Setup}^{\mathcal{H}^{\text{origin}}}(\Sigma^0_{\mathrm{R}_S})$

3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\mathrm{Leak}(m_1))$

4 : $b \leftarrow\!\!\$\ \{0,1\}$

5 : $\hat{\varphi} \leftarrow \mathcal{H}^{\text{origin}}(\mathrm{R}_S(m_b, \Sigma^\star_{\mathrm{R}_S}))$

6 : **return** $\varphi = \hat{\varphi}, b$

Sicherheitsspiel 4: In diesem Sicherheitsspiel wird nur die Darstellung der Absenderin betrachtet.

Die Sicherheitsdefinition ist analog zur vorherigen Definition und lautet:

Definition 10. *[Schutz gegen Herkunftstäuschung] Eine Darstellung* $\Pi = (\mathrm{Gen}, R_S)$ *ist* α, β*-sicher gegen Täuschung einer (durchschnittlichen) Nutzerin* $\mathcal{H}$ *mit* Setup *über Herkunft und Aktion einer Nachricht, gdw. für alle Angreifer* $\mathcal{A}$ *mit* Leak *gilt:*

$$\Pr[\mathit{Origin}^{\mathcal{A},\mathcal{H}^{\mathit{origin}}}_{\mathrm{Setup},\mathrm{Leak},\mathrm{Gen},R_S} = 1, 0 | b = 0] \leq \beta$$
$$\Pr[\mathit{Origin}^{\mathcal{A},\mathcal{H}^{\mathit{origin}}}_{\mathrm{Setup},\mathrm{Leak},\mathrm{Gen},R_S} = 1, 1 | b = 1] \geq \alpha$$

Eingeschränktes Herkunfsspiel

Praktisches Beispiel

Die Herkunft der E-Mail beim gezielten Angriff gegen ägyptische Aktivistinnen in Kapitel 3 (vgl. Abbildung 3.4) ist ein Angriff auf die Darstellung der E-Mail-Adresse.

Wissenschaftliches Beispiel

In Online-Studien sollten Teilnehmerinnen die Herkunft von angezeigten URLs bestimmen [4, 113]. Auf die Darstellung von weiteren Aspekten einer Nachricht oder Webseite wurde in den Experimenten verzichtet. Die menschliche Interpretation und Wahrnehmung von URLs im Allgemeinen wurde dabei untersucht.

Unterschied zum vorherigen Spiel

Es wird nur die Darstellung der Herkunft untersucht und damit der eigentliche Indikator zur Bestimmung der Herkunft.

Eignung

Vorteile	**Nachteile**
Sehr wenig Kontext nötig	Kein kompletter Angriff wird untersucht
Einfaches Experiment	Nicht direkt auf eine
Keine Anwendung nötig	Anwendung übertragbar
Weniger äußere Einflüsse	

Das Spiel und das Experiment sind geeignet, um unabhängig von einer speziellen Anwendung allgemeine Prinzipien zu untersuchen. Für die akademische Forschung eignet es sich zur Untersuchung von allgemeinen Verhalten oder von allgemeinen Verfahren.

Im Folgenden wird der Vergleich zwischen dem Spiel 3 und dem Spiel 4 betrachtet. Der Unterschied ist, dass $\mathcal{H}$ im Spiel 4 eine andere Darstellung angezeigt wird. Dabei ist festzuhalten, dass R_S Teil von R ist und der relevanten Teil zur Darstellung der Herkunft einer Nachricht ist. Daneben umfasst R im Allgemeinen noch die Darstellung vom Inhalt. Dieser Teil wird als R_C bezeichnet. Sei R zusammengesetzt aus R_S und R_C und seien $\mathcal{A}, \mathcal{H}$ und Setup fixiert. Der Zustand $\Sigma^\star_{\mathrm{R}}$ ist zusammengesetzt aus $\Sigma^\star_{\mathrm{R}_S}$ und $\Sigma^\star_{\mathrm{R}_C}$.

Sei m eine Nachricht und gefragt wird nach der Herkunft der Nachricht, wobei $\mathrm{R}(m) = (\mathrm{R}_S(m), \mathrm{R}_C(m))$ und $\Sigma^\star_{\mathrm{R}} = (\Sigma^\star_{\mathrm{R}_S}, \Sigma^\star_{\mathrm{R}_C})$. Im Idealfall leiten Nutzerinnen die Herkunft einer Nachricht nur über die dazu

passende grafische Oberfläche ab. Sei $\hat{\varphi}$ eine Herkunft, dann sollte gelten:

$$\Pr[\mathcal{H}^{\text{origin}}(\mathrm{R}(m), \Sigma^{\star}_{\mathrm{R}}) = \hat{\varphi}] = \Pr[\mathcal{H}^{\text{origin}}(\mathrm{R}_S(m), \Sigma^{\star}_{\mathrm{R}_S}) = \hat{\varphi}]$$

Falls dies aber nicht zutrifft, ist der Inhalt der Nachricht bei der Herkunftsherleitung ebenso relevant. Der Inhalt einer Nachricht kann im Allgemeinen beliebig von der Absenderin, also auch der Angreiferin, gewählt werden und so ist die Ableitung der Herkunft einer Nachricht leicht manipulierbar. Beispielsweise kann die Signatur am Ende einer E-Mail oder das HTML-Layout einer E-Mail von der Angreiferin manipuliert werden.

Wenn der Inhalt der Nachricht die richtige Erkennung der Herkunft fördert, ist dies potentiell gefährlich. Denn es zeigt, dass die Angreiferin den Inhalt der Nachricht nicht gut genug auf die Herkunft abgestimmt hat. Es besteht die Gefahr, dass bei einem besser abgestimmten Inhalt der Nachricht $\mathcal{H}$ von der Darstellung des Inhalts die Bestimmung der Herkunft ableitet und so getäuscht werden kann.

Für die experimentelle Untersuchung der Bestimmung der Herkunft (mittels R_S) ist es darum sinnvoll, den Inhalt einer Nachricht zu ignorieren, weil dieser Stimulus das Ergebnis beeinflussen kann. Die Gesamtwirkung kann in einer konkreten Implementierung und Einbettung in einer Anwendung untersucht werden.

Herkunftsbestimmung

In einer Anwendung sollte die Herkunft einer Nachricht nur durch die explizite Darstellung der Herkunft hergeleitet werden.

Für eine Anwendung ist es sinnvoll, dies bei der Gestaltung der grafischen Oberfläche zu berücksichtigen und zu untersuchen. Dies kann sich je nach Gestaltung der grafischen Oberfläche je nach Anwendung unterscheiden. Bei vielen Anwendungen wird bei einer Auflistung der Nachrichten nur die Herkunft einer Nachricht und ein Textausschnitt über den Inhalt angezeigt. Anwendungen sollten die Bestimmung der Herkunft einer Nachricht in diesem Schritt fokussieren.

Der Fokus ist die Konzeption und Untersuchung zur Darstellung der Herkunft einer Nachricht und im Folgenden nur noch deren Darstellung mit R_S. Dieser Bereich ist dediziert für die Herleitung der Herkunft.

Eine Alternative ist es, die Elemente im Inhalt zur Herleitung der Herkunft kryptographisch abzusichern und die Manipulierbarkeit zu erschweren.

4.5 Einschränkungen

Die vorgestellten Spiele bilden ein Modell und eine neue Perspektive, um Phishing-Angriffe zu betrachten, und haben den Anspruch, konstruktiv die Sicherheit durch die Entwicklung und Analyse von Gegen-

maßnahmen zu verbessern. Allerdings gibt es hier Einschränkungen, welche betrachtet werden sollten. Zunächst ist ein Modell sinnvoll, um einen komplexen und vielseitigen Angriff, welcher sowohl technische als auch mindestens psychologische Aspekte berücksichtigt und im Angriffsfall diese ausnutzt, darzustellen. In dem Modell erfolgt eine Reduzierung auf wesentliche Aspekte und eine Ausklammerung von vielen Faktoren und damit bildet es nicht die komplette Realität ab. Vielmehr ist es ein erster Schritt zu einer konstruktiven Betrachtung und nicht endgültig. Weitere Aspekte können durch andere Varianten von Sicherheitsspielen zusätzlich modelliert werden.

4.5.1 Keine konkrete Umsetzung

Das vorgestellte Modell kann zwar als Grundlage für Experimente betrachtet werden, aber ist zunächst unabhängig davon. Es ermöglicht die Betrachtung und Analyse von abstrakten Verfahren. Ein Prinzip statt einer konkreten Implementierung wird betrachtet. Damit kann das Phänomen auftreten, dass zwar ein prinzipielles Verfahren als vorteilhaft betrachtet wird, aber eine konkrete Implementierung in einer Anwendung angreifbar ist. Dieses Problem ist aus anderen formalen Methoden in der IT-Sicherheit, wie zum Beispiel der modernen Kryptographie, bekannt. Koblitz und Menezes beschreiben als Beispiel einen Angriff gegen eine praktische Implementierung des als sicher bewiesenen Protokolls zur Verknüpfung von zwei Bluetooth-Geräten [71]. In dem Beispiel wurde ein Detail des Protokolls nicht bei der Implementierung berücksichtigt und die Bedeutung dieses Details wurde im Sicherheitsbeweis nicht deutlich. Der mögliche Unterschied zwischen der (theoretischen) Sicherheit in einem bestimmten Modell und der (konkreten) Sicherheit in der Praxis ist eine inhärente Schwäche von Sicherheitsmodellen und betrifft das vorgestellte Modell. Trotzdem ist die Betrachtung von Sicherheit in einem Modell eine akzeptierte und bewährte Methode.

4.5.2 Schwächen außerhalb vom Modell

Formale Modelle bilden nicht die vollständige Realität ab, sondern versuchen die Reduzierung auf das Wesentliche. Die Erfahrungen aus der Kryptographie zeigen, dass die Lücken zwischen Modell und Realität ausgenutzt werden können. In den Modellen wird nicht betrachtet, dass die Angreiferin die Berechnungen beobachten kann, also zum Beispiel die konkrete Berechnung einer verschlüsselten Nachricht mittels RSA. Kocher zeigte, dass bei konkreten Implementierungen von RSA und anderen Algorithmen die Angreiferin die Berechnungszeiten der Implementierung beobachten kann und diese Informationen für einen erfolgreichen Angriff nutzen kann [72]. Diese Angriffsstrategie wurde in den klassischen formalen Modellen der Kryptographie nicht berücksichtigt und ist ein klassisches Beispiel für sogenannte Seitenkanal-Angriffe [71]. Die Sicherheit wird dabei nur innerhalb des Modells mit den Annahmen berücksichtigt. Der vorgeschlagene Weg über ein Modell kann einerseits interpretiert werden, um Seitenkanäle auf menschlicher Ebene zu modellieren. Ebenso ist es möglich, dass es Angriffe außerhalb vom Modell gegen eine konkrete Implementierung geben kann.

Insbesondere sind Spiele, in denen das Verfahren isoliert betrachtet wird, von möglichen Seitenkanälen oder anderen grafischen Oberflächen aus der Anwendung betroffen. Zusätzlich kann gelerntes oder adaptives menschliches Verhalten durch andere Anwendungen die Auswirkung eines Verfahrens beeinflussen. Hierbei wird deutlich, dass die Berücksichtigung von allen Faktoren bei einer Mensch-Computer-Interaktion uferlos ist und vermutlich umfangreicher ist als bei der Nutzung von Kryptographie. Insbesondere dadurch, dass für sichere Mensch-Computer-Interaktionen häufig kryptographischen Verfahren nutzen müssen. Im Bereich der Kryptographie werden diese Einschränkungen im Wesentlichen akzeptiert und sie haben dem Verständnis von Sicherheit in diesem Bereich geholfen.

4.5.3 Keine bedingungslose Sicherheit

Koblitz und Menezes kritisieren insbesondere die Nutzung von Begriffen, wie Theorem, Beweis und beweisbare Sicherheit [71]. Dies ist eine Warnung für den vorgestellten Ansatz. Ein Sicherheitsbeweis kann nicht in diesem Modell erbracht werden. Vielmehr kann das Modell ein Argument für die Nutzung von bestimmten Verfahren sein und die Gründe für die Steigerung der Sicherheit hervorheben. Es wird eine logische und nachvollziehbare Argumentationskette bereitgestellt. Gleichzeitig können Annahmen und Anforderungen explizit genannt werden und ermöglichen einen Vergleich unterschiedlicher Verfahren. Formale Beweise in der Kryptographie basieren meist auf einer Reduktion auf ein (vermutlich) mathematisch schwer lösbares Problem, zum Beispiel einem NP-schweren Problem. Koblitz und Menezes [71] sowie Goldwasser und Kalai [51] kritisieren die neueren kryptographischen Verfahren, die auf relativ unbekannten Problemen basieren. Die Problematik besteht darin, dass diese Probleme relativ unverstanden und nur im Kontext von dem kryptographischen Verfahren interessant sind [51, 71]. Die hervorgehobene Problematik dabei ist, dass eine Tautologie besteht, indem die Sicherheit eines Verfahrens auf einem Problem besteht und die Schwierigkeit des Problems auf der Sicherheit eines Verfahrens basiert [51, 71]. Dies ist eine Warnung bei der Konstruktion eines Verfahrens in diesem Modell. Die Sicherheit eines Verfahrens sollte grundsätzlich auf Problemen und Erkenntnissen allgemeinerer Natur basieren. Sehr spezielle Annahmen, welche zu uninteressant zur alleinigen Untersuchung sind, sind zu hinterfragen und sind ungeeignete Annahmen. Gleichzeitig besteht die Gefahr, dass eben eine allgemeine Erkenntnis über die menschliche Kognition nicht exakt zu einem speziellen Verfahren passt und damit ungenau ist. Es deutet sich somit ein Spannungsfeld an, welches an einem konkreten Verfahren genauer betrachtet wird und abgewogen werden muss. Dies zeigt, dass nur Argumente für die Sicherheit eines Verfahrens gefunden werden können, aber es keine bedingungslose Wahrheit und damit keinen Beweis für die Sicherheit gibt.

Die Sicherheitsdefinitionen sind Abschätzungen der Eintrittswahrscheinlichkeiten von Ereignissen. Durch die Vereinfachung der Sicherheitsspiele kann es zu Abweichungen und Veränderungen der Eintrittswahrscheinlichkeiten kommen.

4.6 Zusammenfassung

In diesem Kapitel wurden verschiedene Sicherheitsspiele betrachtet und verglichen. Der Angriff wurde unterschieden zwischen der Täuschung über die Herkunft und der gewünschten Aktion. Außerdem wurde die Darstellung einer Nachricht in die Herkunft und den Inhalt getrennt.

Ein komplexer Angriff wurde unterteilt und damit der Umfang bei der Untersuchung reduziert. Gleichzeitig ermöglicht dies eine Fokussierung und Reduzierung auf die wichtigen Sicherheitsmechanismen.

Im ersten Schritt wurde das Sicherheitsspiel durch die Fokussierung auf die Herkunft einer Nachricht vereinfacht. Im zweiten Schritt wird das Sicherheitsspiel durch die Fokussierung der Darstellung der Herkunft vereinfacht. Im Ergebnis bedeutet dies, wenn ein Angriff im ersten Sicherheitsspiel möglich ist, dann ist ein Angriff im zweiten und dritten Spiel möglich. Die Kontraposition ermöglicht die Aussage, dass wenn das dritte Spiel nicht angegriffen werden kann, dann nicht das erste und deutlich komplexere Sicherheitsspiel.

$\text{Phishing}^{\mathcal{A},\mathcal{H}^{\text{origin \& sudo}}}_{\text{Setup,Leak,Gen,R}}$

1 : $\Sigma_{\text{R}}^{0} \leftarrow \text{Gen}()$
2 : $\Sigma_{\text{R}}^{\star}, m_1, \varphi, \gamma \leftarrow \text{Setup}^{\mathcal{H}^{\text{origin \& sudo}}}(\Sigma_{\text{R}}^{0})$
3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{\varphi}, \hat{\gamma} \leftarrow \mathcal{H}^{\text{origin \& sudo}}(\text{R}(m_b, \Sigma_{\text{R}}^{\star}))$
6 : $\hat{b} \leftarrow \varphi = \hat{\varphi} \wedge \gamma = \hat{\gamma}$
7 : **return** $\hat{b}, b$

$\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}}$

1 : $\Sigma_{\text{R}}^{0} \leftarrow \text{Gen}()$
2 : $\Sigma_{\text{R}}^{\star}, m_1, \varphi \leftarrow \text{Setup}^{\mathcal{H}^{\text{origin}}}(\Sigma_{\text{R}}^{0})$
3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{\varphi} \leftarrow \mathcal{H}^{\text{origin}}(\text{R}(m_b, \Sigma_{\text{R}}^{\star}))$
6 : **return** $\varphi = \hat{\varphi}, b$

Trotzdem erfordern die verschiedenen Sicherheitsspiele 2, 3 und 4 immer noch eine komplexe Simulation der Nutzerinnen und diese Ausgestaltung ist eine große Herausforderung. Das Ziel der vorgeschlagenen Methode ist eine Fokussierung auf Sicherheitsmechanismen und die Erläuterung, warum diese die Sicherheit erhöhen und stärken. Aus den Sicherheitsspielen lassen sich bereits einige Fehlerquellen bei der Darstellung einer Nachricht ableiten. Das Sicherheitsspiel 2 umfasst einen großen Kontext. Insbesondere die Aktion und die legitime Herkunft müssen zusammenpassen. Ansonsten beeinflusst es die Sicherheitsanalyse. Insbesondere die empirischen Ergebnisse bei Experimenten, welche von diesem Sicherheitsspiel angelehnt sind, sind stark abhängig von der Aktion und der Herkunft. Die Verallgemeinerung von den Experimenten ist kaum möglich. Diese Sicherheitsspiele sind insbesondere für Organisationen oder bestimmte Personengruppen sinnvoll, um die Sicherheit der Anwendung in dem genutzten Kontext zu evaluieren.

Das Sicherheitsspiel 3 untersucht die Bestimmung einer Herkunft in einer Anwendung und ist darum besonders von der Anwendung abhängig. Eine Sicherheitsanalyse basierend auf empirischen Ergebnissen ist nur bedingt auf andere Anwendungen übertragbar. Insbesondere Designentscheidungen und Interakti-

onskonzepte unterscheiden sich je nach Anwendung und Plattform. [3] Die Verallgemeinerung auf andere Plattformen und Anwendungen ist damit nur sehr eingeschränkt möglich. Diese Art von Untersuchungen ist damit insbesondere für die Hersteller von Anwendungen von Relevanz und einfach möglich. Vergleichbare Studien werden oftmals im Entwicklungsprozess bereits durchgeführt, um die Nutzbarkeit zu studieren.

Das Sicherheitsspiel 4 beschränkt sich bei der Bestimmung der Herkunft auf bestimmte Bereiche und ist unabhängig von einer konkreten Anwendung. Es werden damit Prinzipien und abstrakte Verfahren untersucht und nach Möglichkeit kann die Sicherheitseinschätzung basierend auf kognitive Forschungserkenntnisse bestärkt werden. Diese Ebene ist insbesondere für die akademische Forschung interessant, weil diese unabhängig von Anwendungen sind und eher die Grundlagen für die Sicherheit untersuchen.

Die Sicherheitsspiele 3 und 4 untersuchen die Ableitung der Herkunft. Ähnliche Sicherheitsspiele können für Aktionen ebenso entworfen werden. In der Arbeit wird stattdessen auf die Herkunft einer Nachricht fokussiert und die Formalisierung konstruktiv genutzt.

4.6.1 Verkettung von Sicherheitsspielen

In diesem Kapitel wurde der Phishing-Angriff schrittweise vereinfacht und in mehrere Sicherheitsspiele unterteilt. Dies verdeutlicht die unterschiedlichen Aspekte einen Angriffs und ermöglicht die Konzentration auf einen Wirkungsmechanismus, um die Sicherheit gegen einen Angriff zu erhöhen. Damit können die Wirksamkeit und die Bedeutung von einem Darstellungsbereich genauer untersucht und der Kern der Sicherheitserhöhung hervorgehoben werden. Die Konzentration auf den Kern der Sicherheit ermöglicht einen konstruktiven Ansatz, um den Schutz gegen Angriffe und vor Fehlern von Nutzerinnen in einer Anwendung zu erhöhen. Ein Konzept von einem Darstellungsverfahren wird im späteren Verlauf so untersucht.

In einigen Feldstudien wurde die Anfälligkeit von Nutzerinnen gegen Phishing-Angriffe untersucht und ein typischer Phishing-Angriff wird simuliert [27, 35, 47, 52]. Häufig erhalten die Personen eine fingierte E-Mail mit einem Link zu einer fingierten Webseite. Auf dieser Webseite werden die Personen aufgefordert, ihr Passwort einzugeben. Ein erfolgreicher Phishing-Angriff ist meist die Eingabe von einem Passwort auf der Webseite oder in manchen Fällen der Besuch der Webseite. In diesen Experimenten wird die genutzte Software vernachlässigt oder die Nutzung der gleichen Software unter den Teilnehmenden angenommen. Im Gegensatz zu der Analyse durch Sicherheitsspiele wird der Kern der Sicherheit einer Anwendung nicht untersucht. Dadurch sind die Ergebnisse aus diesen Sicherheitsspielen nicht direkt vergleichbar mit den Ergebnissen aus diesen Studien.

In der Praxis findet ein Phishing-Angriff häufig über verschiedene Anwendungen, zum Beispiel E-Mail-Anwendung und Web-Browser, statt. Dies ist so in einem Sicherheitsspiel nicht vorgesehen. Das Angriffsszenario ist eine Verkettung von Darstellungen und Anwendungen.

[3] Beispielsweise unterscheiden sich die Designrichtlinien von Google und Apple für ihre jeweiligen Smartphone-Plattformen deutlich.

Für die Sicherheitsdefinitionen bedeutet das, dass ein praktischer Angriff eine Abfolge von Erfolgen in Sicherheitsspielen unterschiedlicher Darstellungen ist. Die Sicherheitsspiele werden über die Anwendungen und ggf. innerhalb der Anwendungen miteinander verkettet. Hierfür eigenen sich besonders die Sicherheitsspiele mit einer Aktion γ. Innerhalb der Kette ist die Aktion der Start eines neuen Sicherheitsspiels in einer anderen Darstellung oder Anwendung. Erst im letzten Sicherheitsspiel der Kette wird durch die Aktion das Ziel der Angreiferin erreicht. Die Verkettung von Sicherheitsspielen kann sowohl innerhalb einer Anwendung erfolgen oder über Anwendungen hinaus. Die Verkettung von Sicherheitsspielen erfolgt parallel, wenn mehrere Darstellungen gleichzeitig sichtbar sind, zum Beispiel die Herkunft einer Nachricht und der Inhalt einer Nachricht, oder hintereinander, wenn eine Aktion aus einem Sicherheitsspiel wieder eine neue Darstellung öffnet. Innerhalb einer Anwendung wird beispielsweise eine Nachricht im Postfacheingang ausgewählt und die ausgewählte Nachricht wird im Detail dargestellt. Der Klick auf einen Link in einer Nachricht startet ein neues Sicherheitsspiel in einem Browser.

In diesem Beispiel eines klassischen Angriffs zur Preisgabe des Passworts gibt es folgende Darstellungen und Aktionen:

1. Die Nutzerinnen sehen eine reduzierte Darstellung der Nachricht innerhalb des Posteingangs innerhalb der E-Mail-Anwendung. Die Zielaktion ist das Öffnen der E-Mail.
2. Die Nutzerinnen sehen eine Darstellung der E-Mail innerhalb der E-Mail-Anwendung. Die Zielaktion ist der Klick auf den Link.
3. Die Nutzerinnen besuchen die verlinkte Webseite im Browser. Das Ziel ist die Eingabe vom Passwort.

Diese Betrachtung verdeutlicht die Hürden der Angreiferin. Es muss jedes Sicherheitsspiel gewonnen werden, damit der Angriff erfolgreich ist, und bei Abbruch innerhalb dieser Kette sind die Nutzerinnen immer noch sicher geblieben, obwohl eine Gefährdung vorlag. Allerdings können die Spiele voneinander abhängen. Die Herkunft kann nur in der ersten Darstellung berücksichtigt werden. In den weiteren Darstellungen werden diese nicht mehr berücksichtigt.

Beispielsweise wird in einem klassischen Phishing-Angriff vorgetäuscht, dass die E-Mail von PayPal verfasst wurde. Nutzerinnen werden aufgefordert die PayPal-Webseite zu besuchen und dort das Passwort für das PayPal-Nutzerkonto einzugeben. In diesem Fall können Nutzerinnen im Posteingang (also in einer Liste von E-Mails) zur Entscheidung kommen, dass die E-Mail von PayPal stammt. Beim Lesen der E-Mail beachten sie dann nicht mehr die Darstellung der Herkunft. Wenn sie dann auf den Link in der E-Mail klicken und die Webseite besuchen, dann könnten sie nicht mehr die URL beachten, weil angenommen wird, dass die Herkunft PayPal ist. In diesem beispielhaften Szenario wird die Herkunft dann nur im Posteingang bestimmt und dann in den nächsten Spielen weiterhin angenommen. Damit ist die Herleitung der Herkunft zwischen den Sicherheitsspielen nicht unabhängig.

Anderseits kann die Gesamterfolgswahrscheinlichkeit die Multiplikation der einzelnen Erfolgswahrscheinlichkeiten möglich sein, falls alle Sicherheitsspiele von der Angreiferin gewonnen werden müssen und die einzelnen Sicherheitsspiele (stochastisch) unabhängig sind. Das ist aber eine offene Forschungsfrage.

Im Folgenden wird dieser ganzheitliche Ansatz nicht weiter verfolgt und stattdessen wird die Konstruktion von einem sicheren Verfahren und die Betrachtung von E-Mail-Adressen in diesem Modell erfolgt. Beim ganzheitlichen Ansatz ist eine wichtige offene Forschungsfrage die Abhängigkeit zwischen den einzelnen Sicherheitsspielen in der Folge der Spiele.

Praxis und Sicherheitsspiele

Der Erfolg eines praktischen Angriffs ist die Verkettung von Erfolgen in verschiedenen Sicherheitsspielen über mehrere Anwendungen und Darstellungen hinweg.

Kapitel 5

Gefährliche Zeichenketten

Bevor im nächsten Kapitel die Formalisierung von Phishing zur Konstruktion von einem Verfahren genutzt wird, werden in diesem Kapitel die Gefahren bei der Darstellung von E-Mail-Adressen illustriert. Damit wird zunächst die Anwendbarkeit demonstriert und gleichzeitig zeigt es die Notwendigkeit eines sicheren Verfahrens. Das zuvor vorgestellte Sicherheitsmodell ermöglicht die Aufstellung allgemeiner naheliegender Heuristiken[1], welche potentiell eine Gefahr aufzeigen und die Sicherheit eines Verfahrens in Frage stellen. Als illustrierendes Beispiel wird die typische Darstellung der E-Mail-Adresse als eine Zeichenkette verwendet. Mittels dieser Zeichenkette bestimmt der Mensch die Herkunft einer E-Mail. In diesem Kapitel wird ein Überblick über mögliche Gefahrenquellen bei dieser Darstellung aufgezeigt und um Erkenntnisse aus der Kognitionswissenschaft ergänzt.

5.1 Identische Täuschungen

In diesem Abschnitt wird die Möglichkeit eines Angriffs, in dem die Herkunft einer Nachricht exakt gefälscht wird, betrachtet. In diesem Fall kann der Mensch die Fälschung nicht erkennen und ein erfolgreicher Angriff ist naheliegend.

Unsicherheitsheuristik: Identische Kopie

Eine Darstellung $\Pi = (\mathrm{R}_S, Gen)$ ist unsicher, wenn eine Angreiferin $\mathcal{A}$ existiert, die im Sicherheitsspiel $\mathrm{Origin}^{\mathcal{A},\mathcal{H}^{\mathrm{origin}}}_{\mathrm{Setup,Leak,Gen,R}_S}$ mit der Nachricht m_0 eine Nachricht m_1 erzeugen kann und folgende Eigenschaft erfüllt ist:

$$\mathrm{R}_S(m_0, \Sigma^{\star}_{\mathrm{R}_S}) = \mathrm{R}_S(m_1, \Sigma^{\star}_{\mathrm{R}_S})$$

[1] Die Heuristiken wirken sehr naheliegend und trivial. Literatur, in denen dies explizit genannt wird, sind mir aber nicht bekannt. In der Vielzahl der Literatur zum Thema Phishing mag es diese aber geben.

Begründung:

Seien m_0, m_1 die Nachrichten aus dem Sicherheitsspiel $\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}_S}$ mit $r_0 = \text{R}_S(m_0, \Sigma^{\star}_{\text{R}_S})$ und $r_1 = \text{R}_S(m_1, \Sigma^{\star}_{\text{R}_S})$ mit $r_1 = r_0$. Dann kann der Mensch beide Nachrichten nicht unterscheiden und die Wahrscheinlichkeitsverteilung der Antwort ist unabhängig von der Präsenz eines Angriffs (also formal im Spiel 3 unabhängig vom Bit b). Es gilt somit:

$$\Pr[\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}_S} = 1, 0] = \Pr[\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}_S} = 1, 1]$$

$\mathcal{H}$ leitet bei den Darstellungen von m_0, m_1 mit einer gleichen Wahrscheinlichkeit die legitime Identität ab und die Darstellung erhöht nicht die Sicherheit.

Dies folgt direkt aus der Definition von bedingten Wahrscheinlichkeiten und das $\mathcal{H}$ stochastisch unabhängig von b ist.

Damit gilt $\alpha = \beta$ und somit ist das Verfahren nach der Definition von Brauchbarkeit (vgl. Definition 4.1.3) nicht brauchbar als Sicherheitsmechanismus.

Diese allgemeine Heuristik ist eine naheliegende Herausforderung und dennoch werden potentielle Gefahrenstellen in der E-Mail aufgezeigt.

5.1.1 Bekannte Fallstricke

Im E-Mail-Protokoll wird die Darstellung der Herkunft vom `FROM`-Feld abgeleitet. In das Textfeld kann eine beliebige Zeichenkette eingetragen werden und das Feld ist zunächst nicht kryptographisch gesichert. In Abschnitt 3.6.1. im RFC 5322 [111] wird zur Sicherheit über die Herkunft Folgendes geschrieben:

> In all cases, the From: field SHOULD NOT contain any mailbox that does not belong to the author(s) of the message.

Die Bitte, nur die echte Herkunft anzugeben, zeugt von einer gewissen Naivität, die anscheinend bösartige Angriffe ignoriert. Erst mit der Einführung weiterer RFCs wird diese Schwachstelle ausgebessert. In der Praxis von E-Mail-Anwendungen wird die Darstellung der Herkunft auf das `FROM`-Feld reduziert und die menschliche Entscheidung berücksichtigt diese Darstellung bei der Bestimmung der Herkunft. Wenn in dem Sicherheitsspiel 3 $\mathcal{A}$ durch Leak das `FROM`-Feld von einer legitimen Nachricht kennt, dann kann $\mathcal{A}$ in seiner Nachricht das identische `FROM`-Feld angeben. Ohne weitere Schutzmaßnahmen wird die identische Herkunft dargestellt.

Eine Verhinderung dieses trivialen Angriffs erfordert eine kryptographische Absicherung oder im Notfall eine Überprüfung der Legitimität des ausgehenden Servers. Letzteres obliegt aber dem eingehenden Server. Die Darstellung einer Herkunft sollte darum auf einer kryptographischen Funktion zur Verifikation der Herkunft aufbauen. Diese Funktion wird künftig mit `vrf` bezeichnet und hat als Eingabe eine Nachricht.

Anforderung an `vrf`

Es gibt zwei Anforderungen an $\texttt{vrf}$:

1. Seien m_0, m_1 zwei Nachrichten unterschiedlicher Herkunft. Dann gilt: $\texttt{vrf}(m_0) \neq \texttt{vrf}(m_1)$
2. Seien m_0, m_1 Nachrichten mit identischer Herkunft. Dann gilt: $\texttt{vrf}(m_0) = \texttt{vrf}(m_1)$

Im Kontext der E-Mail kann dies mittels digitaler Signaturen der Protokolle DKIM, PGP oder S/MIME realisiert werden. Bei der Konzeption, Umsetzung und Nutzung dieser Protokolle gibt es weitere Fallstricke. [30, 55, 65, 84, 92, 93, 94, 118, 138] Shen et al. haben die Spezifikation der E-Mail analysiert und folgende Fallstricke gefunden [65]:

1. Verwirrung mit den Feldern `FROM` und `SENDER`
2. Mehrfache Vorkommen vom `FROM`-Feld
3. Verschiedene E-Mail-Adressen in einem Feld
4. Verwirrung durch den Anzeigenamen des `FROM`-Feldes

RFC 5322 [111] spezifiziert neben dem `FROM`-Feld ein `SENDER`-Feld. Der genannte Anwendungsfall ist, dass eine Person (hier Alex) für eine andere dritte Person (hier Vic) eine E-Mail verschickt. In diesem Fall wird die E-Mail-Adresse von Vic im `FROM`-Feld und die tatsächlich verwendete Adresse, also die von Alex, in das `SENDER`-Feld eingetragen. Die Beziehung zwischen den Adressen im `FROM`-Feld und im `SENDER`-Feld ist unklar und es ist nicht gesichert, dass `SENDER` für `FROM` E-Mails versenden darf. Das gilt insbesondere, wenn beide eine unterschiedliche Domain haben. Die Gefahr ist, dass eine Anwendung das `SENDER`-Feld ignoriert und nicht darstellt. Andererseits ist die Darstellung beider Felder für Menschen ohne technischen Hintergrund eine Herausforderung und benötigt Unterstützung oder Schulungen für diesen Sonderfall.

In diesem Fall kann $\mathcal{A}$ wieder das legitime `FROM`-Feld kopieren und seine eigene Adresse ins `SENDER`-Feld eintragen. Die Gefahr besteht, dass weiterhin nur das `FROM`-Feld dargestellt wird und somit die Darstellungen identisch sind. Im Protokoll gibt es keinerlei Hinweise, wie die Herkunft einer E-Mail dargestellt werden sollte und Empfehlungen mit der Darstellung fehlen.

Protokolle für Menschen

Das E-Mail-Protokoll ist zwar für Menschen lesbar, aber ebenso sollte die Darstellung für Menschen berücksichtigt werden.

Shen et al. weisen außerdem auf Gefahren hin, wenn das `FROM`-Feld mehrfach in der E-Mail vorkommt [65]. Im schlimmsten Fall nutzt der Server bei der Überprüfung der Herkunft eine anderes `FROM`-Feld als die Anwendung zur Darstellung der Herkunft. Die Annahmen über die Herkunft einer E-Mail unterscheiden sich damit zwischen Server und Anwendung. Damit passt die Sicherheitsanalyse vom Server nicht mehr zu der dargestellten Herkunft für die Nutzerin. In diesem Fall besteht weiterhin die Gefahr einer iden-

tischen Darstellung. Beispielsweise kann ein `FROM`-Feld die E-Mail-Adresse `service@paypal.com` und ein weiteres `FROM`-Feld die E-Mail-Adresse `adversary@bad-example.com` sein. Wenn eine Nutzerin PayPal kennt und die E-Mail-Adresse `service@paypal.com` angezeigt bekommt und der Server überprüft hat, dass die E-Mail von `adversary@bad-example.com` kommt, dann kann die Nutzerin getäuscht werden und diese Täuschung nicht ohne die Betrachtung alle Meta-Informationen erkennen.

Eine weitere beschriebene Möglichkeit ist das Ausnutzen von Unicode-Steuerzeichen. Beispielsweise kann ein Text in einer anderen Richtung als der Schreibrichtung dargestellt werden, das heißt `lapyap` kann per Unicode Zeichen als `paypal` dargestellt werden und damit können Menschen getäuscht werden [65]. Dies ist beispielsweise sinnvoll, wenn in einem arabischen Text Eigennamen mit lateinischen Buchstaben vorkommen und diese dann von links-nach-rechts geschrieben werden, aber später von rechts-nach-links gelesen werden.

Eine weitere Gefahr besteht durch den Aufbau von E-Mail-Adressen. Im RFC2822 [110] wird neben der E-Mail-Adresse noch ein Anzeigename spezifiziert. Dieser ist ein beliebiger Text und wird nicht zwangsläufig vom Server geprüft. Die Verschleierung der tatsächlichen Herkunft der Nachricht wird ermöglicht. Eine Angreiferin $\mathcal{A}$ kann den gleichen Namen oder die tatsächliche Adresse als eigenen Namen verwenden und so eine identische Darstellung im Vergleich zur legitimen Herkunft erzielen. Mittlerweile stellen gängige E-Mail-Anwendungen, wie AppleMail, Thunderbird oder Outlook, Anzeigenamen mit einem @-Zeichen nicht mehr dar oder zeigen sowohl den Anzeigenamen als auch die E-Mail-Adresse an.

Eine identische Täuschung basiert auf Unzulänglichkeiten im Protokoll bzw. deren Implementierungen und ist für Menschen nur schwer erkennbar. Eine Angriffserkennung ist teilweise bei der Betrachtung und dem Vergleich der E-Mail als Textdatei möglich, aber oftmals nicht beim (ersten) Blick auf die Darstellung in einer Anwendung. Beim Entwurf und der Implementierung muss diese Gefahr berücksichtigt werden.

5.2 Fehlerhafte Wahrnehmung

Im vorherigen Abschnitt wurde gezeigt, dass eine Angreiferin versucht, eine identische Darstellung zu erzeugen. Wenn dies nicht möglich ist, gibt es noch weitere Gefahrenquellen. Eine weitere Gefahrenquelle ist, dass zwar keine identische Darstellung von der Angreiferin erzeugt wird, aber Nutzerinnen den Unterschied nicht bemerken. Zunächst ist eine Beschreibung von einem möglichem kognitiven Fallstrick nötig und im Anschluss wird der Bezug zu einem Angriff herausgearbeitet.

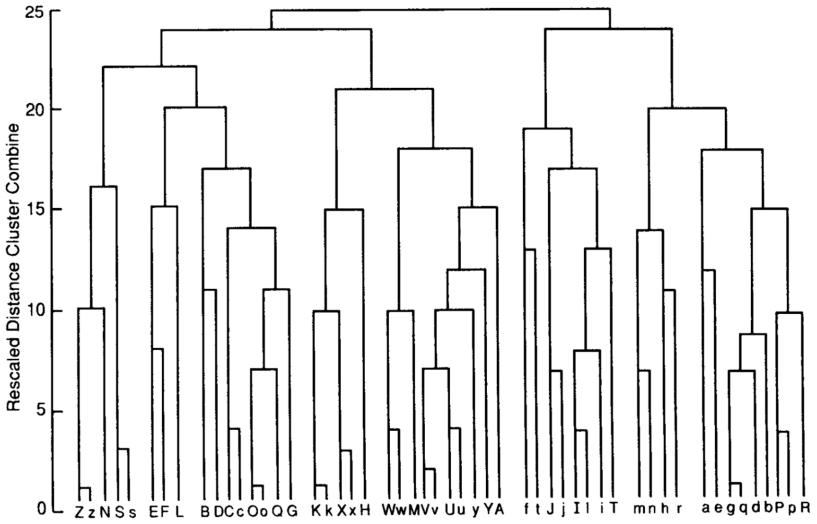

Figure 2. The WPGMA cluster analysis.

Abbildung 5.1: Abstandscluster nach einer Umfrage von Boles und Clifford [19]

5.2.1 Buchstaben-Substitution

Aus den Kognitionswissenschaften und der Psychologie sind verschiedene Phänomene bei der Wahrnehmung von Zeichenketten bekannt. Ein bekannter Effekt ist, dass der Austausch von einem Buchstaben durch einen anderen Buchstaben unterschiedlich wahrgenommen und erkannt wird. Im Experiment von Healy [57] wurden in einem Text bei Wörtern ein `s`, `c`, `k` oder `p` vor einem `a` entfernt und Studienteilnehmende markierten Fehler im Text (*proof-reading task*). Dabei wurden signifikant weniger Fehler bei `s` und `c` Entfernungen als bei `k` und `p` erkannt und markiert.

Eine Hypothese zur Beschreibung der unterschiedlichen Erkennung von Fehlern ist die Form der Buchstaben bzw. die Veränderung der Form des gesamten Wortes. Mueller und Weidemann untersuchten die Wahrnehmung von einzelnen Buchstaben [91]. Studienteilnehmenden wurde zunächst ein Buchstabe für kurze Zeit gezeigt und im Anschluss wurde der gleiche Buchstabe und ein weiterer Buchstabe angezeigt. Die Teilnehmenden sollten schnell angeben, welcher der beiden vorher dargestellt wurde. Diese Art von Experiment wurde bereits in der Vergangenheit häufig durchgeführt und ist als *2-AFC perceptual letter identification task* bekannt [91]. In diesem Experiment wurde deutlich, dass sowohl die Fehlerraten bei Buchstaben als auch die Antwortzeiten sich deutlich unterschieden. In der Studie wurden nur Großbuchstaben verwendet und darum auf E-Mail-Adressen nicht übertragbar.

Boles und Clifford haben Studienteilnehmende gebeten, die Ähnlichkeit zwischen zwei Buchstaben zu bewerten und haben daraus ein Abstandscluster konstruiert [19]. Abbildung 5.1 stellt deren berechnetes Cluster dar. Beispielsweise werden `g` und `q` als sehr ähnlich bewertet.

Der direkte Bezug zwischen den unterschiedlichen Experimenten ist nicht offensichtlich, weil die Experimente in unterschiedlichen Kontexten erfolgten und sich im Studiendesign unterscheiden. Gemein ist diesen Studien aber, dass Buchstaben ähnlich sein können und eine fehlerhafte Wahrnehmung möglich ist.

Nicht Erkennung einer Substitution

Kognitive Experimente zeigen, dass Menschen einen Austausch eines Buchstabe nicht immer erkennen.

Diese Eigenschaft der menschlichen Wahrnehmung unterstützt im Alltag, Texte trotz einer falschen Rechtschreibung lesen und verarbeiten zu können. Bei einer E-Mail-Adresse oder einer URL hingegen kann eine solche fehlerhafte Wahrnehmung einen erfolgreichen Angriff bedeuten. Die Angreiferin versucht, eine solche fehlerhafte Wahrnehmung wie in den obigen Beispielen zu erzwingen.

Eine direkte Übertragung auf die Sicherheitsanalyse der Darstellung der Herkunft mittels einer E-Mail-Adresse mit einer Erfolgswahrscheinlichkeit ist nicht möglich. Es zeigt die Schwierigkeiten bei einer nachträglichen Sicherheitsanalyse und verdeutlicht den Bedarf weiterer stark vereinfachter Heuristiken. Insbesondere bei E-Mail-Adressen und URLs sind Angriffe mittels dem Austausch eines Buchstabens innerhalb einer Domain plausibel. Simpson et al. untersuchten diese Angriffe in der Praxis [125]. Sie haben dafür 12 mögliche Substitutionen, wie zum Beispiel die Ersetzung von einem `g` mit einem `q`, betrachtet und alle registrierten Domains von der `.com`-Zone zwischen 2009 und 2019 betrachtet. Für $95,1\%$ aller betrachteten Firmen (269.759 Firmen) war eine visuell sehr ähnliche Domain registriert.

5.2.2 Wahrnehmungskollision

Die obigen Experimente zeigen, dass im Alltag Menschen den Unterschied zwischen zwei Buchstaben nicht erkennen können. Für eine abstrakte allgemeine Heuristik bietet sich eine Vereinfachung abseits von den unterschiedlichen Studiendesigns an.

Zunächst wird eine Anfrage an $\mathcal{H}$ mit equal definiert. Diese Anfrage hat zwei Eingaben und $\mathcal{H}$ entscheidet, ob diese gleich sind (Rückgabewert 1) oder nicht (Rückgabewert 0). Diese Anfrage simuliert einen möglichen Wahrnehmungsfehler. Sie orientiert sich an der Aufgabe im Experiment von Mueller und Weidemann [91]. Mit dieser Anfrage wird die Plausibilität von einer Verwechselung gezeigt und es wird auf eine genaue Angabe der Wahrscheinlichkeiten verzichtet. Im Bezug auf die Buchstabensubstitution ist es plausibel, dass Menschen ein `g` und ein `q` verwechseln und somit zwei Zeichenketten, bei denen nur ein `g` mit einem `q` ausgetauscht wird, als gleich wahrnehmen; also $\mathcal{H}^{\texttt{equal}}(\texttt{google, gooqle}) = 1$.

Die Herausforderung für die Angreiferin ist es, die gleiche Wahrnehmung von unterschiedlichen Elementen zu erzwingen. Die Gleichheit zwischen zwei Darstellungen kann als eine Kollision im Darstellungsraum bei der menschlichen Betrachtung interpretiert werden. Der Begriff der Kollision ist aus der Kryptographie bekannt.

Katz und Lindell geben drei mögliche Arten einer Kollision einer Hashfunktion [66] an:

1. Kollisionsresistenz: Finde zwei unterschiedliche Elemente, welche auf das gleiche Element abgebildet werden.
2. Zweite Urbild-Resistenz: Gegeben sei ein Element und es soll ein zweites Element (Urbild) gefunden werden, welches auf dasselbe Element abbildet.
3. Urbild-Resistenz: Finde ein Element (Urbild), welches auf ein bestimmtes Element abgebildet wird.

Eine visuelle Kollision über den Darstellungsraum wird mit folgenden Unterschieden definiert. Die visuelle Kollision ist nur über den Darstellungsraum definiert und der Darstellungsraum wird mit $\mathcal{L}$ bezeichnet. In den Kollisionsspielen wird nur die Möglichkeit, ob zwei Darstellungen von einem Menschen als gleich bewertet werden, betrachtet. Es werden die technischen Aspekte der Nachrichtenerstellung vernachlässigt und die Darstellung steht im Fokus. Im Gegensatz zu den kryptographischen Definitionen von Kollisionen muss berücksichtigt werden, dass der Gleichheitstest vom Menschen kommt und dieser muss sich ehrlich verhalten. Aus diesem Grund wird der Mensch zufällig mit der gleichen oder zwei verschiedenen Darstellungen befragt.

Ähnlich zur Kryptographie gibt es verschiedene Arten von Kollisionen und zwar wie folgt:

1. Visuelle Kollision (VisCol): Eine Angreiferin kann zwei unterschiedliche Darstellungen finden, welche vom Menschen als gleich wahrgenommen werden.
2. Ausgewählte visuelle Kollision (SelVisCol): Eine Angreiferin kann zu einer gegebenen Darstellung eine andere Darstellung finden und beide werden vom Menschen als gleich wahrgenommen.
3. Zufällige visuelle Kollision (RanVisCol): Zwei zufällig ausgewählte Darstellungen werden vom Menschen als gleich wahrgenommen.

Diese Kollisionen können als kognitiven Spiele dargestellt werden und zwar wie folgt:

$\text{VisCol}_{\mathcal{L}}^{\mathcal{A},\mathcal{H}^{\text{equal}}}$

1 : $s_0, s_1 \leftarrow \mathcal{A}(\mathcal{L})$
2 : **if** $s_0 = s_1$:
3 : **return** $0, 0$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{b} \leftarrow \mathcal{H}^{\text{equal}}(s_1, s_b)$
6 : **return** $\hat{b}, b$

$\text{SelVisCol}_{\mathcal{L},s_1}^{\mathcal{A},\mathcal{H}^{\text{equal}}}$

1 : $s_0 \leftarrow \mathcal{A}(s_1, \mathcal{L} \setminus s_1)$
2 : $b \leftarrow\!\!\$\ \{0,1\}$
3 : $\hat{b} \leftarrow \mathcal{H}^{\text{equal}}(s_1, s_b)$
4 : **return** $\hat{b}, b$

$\text{RanVisCol}_{\mathcal{L}}^{\mathcal{H}^{\text{equal}}}$

1 : $s_1 \leftarrow\!\!\$\ \mathcal{L}$
2 : $s_0 \leftarrow\!\!\$\ \mathcal{L} \setminus \{s_1\}$
3 : $b \leftarrow\!\!\$\ \{0,1\}$
4 : $\hat{b} \leftarrow \mathcal{H}^{\text{equal}}(s_1, s_b)$
5 : **return** $\hat{b}, b$

Kognitionsspiel 1: Es werden die visuelle Kollisionen als kognitive Spiele (1 bis 3) dargestellt. Die kognitive Spiele sind die Verbindung zwischen kognitiven Experimenten und Sicherheitsspielen.

Eine Kollision liegt vor, wenn $\hat{b} = 1$ und $b = 0$ gilt.

Die erste Definition einer Kollision ähnelt einem kontrollierten Experiment, in dem $\mathcal{A}$ durch die Experimentleitung simuliert wird und diese alle Stimuli im Experiment kontrolliert. Die zweite Definition ist von einer Darstellung abgeleitet, welche von $\mathcal{A}$ beeinflusst ist und wird unter anderem bei der obigen Unsicherheitsheuristik benötigt. In dem Spiel zu dieser Definition wird lediglich die Darstellung betrachtet und nicht die technischen Herausforderungen, um eine solche Nachricht zu erzeugen. Das Ignorieren von technischen Details, wie zum Beispiel die Domain-Registrierung und ähnlichem ist sinnvoll, wenn im Wesentlichen die Darstellung in der Endanwendung untersucht wird. In diesem Fall kann davon ausgegangen werden, dass eine motivierte Angriffspartei die Kapazitäten, Fähigkeiten und Möglichkeiten zum Lösen solcher technischen Details besitzt.

Unsicherheitsheuristik: Wahrnehmungskollision

Sei $\Pi = (\mathrm{R}_S, Gen)$ ein Darstellungsverfahren, $\mathcal{L}$ der Darstellungsraum von R_S und r_1 die Herkunftsdarstellung einer legitimen Nachricht m_1 von Setup. Dann ist Π unsicher, wenn eine Angreiferin $\mathcal{A}$ Folgendes kann:

1. Es ist plausibel, dass $\mathcal{A}$ eine visuelle Kollision für r_1 erzeugen kann. Sei r_0 die kollidierende Darstellung zu r_1.
2. Es ist technisch plausibel, dass $\mathcal{A}$ eine Nachricht m_0 mit r_0 als Darstellung der Herkunft erzeugen kann.

Wenn eine Angreiferin eine Nachricht m_0 mit einer visuellen Kollision im Bezug auf die Herkunft zur legitimen Nachricht erzeugen kann, ist es plausibel, dass Nutzerinnen über die Herkunft getäuscht werden. Ein erfolgreicher Angriff ist plausibel. Durch die starke Vereinfachung bei der visuellen Kollision kann keine direkte Wahrscheinlichkeit angegeben werden. Dafür ist die Nutzung der Heuristik sehr einfach und ohne die Durchführung von eigenen Experimenten möglich. Gleichzeitig können weitere Experimente die Verwechselungsgefahr untersuchen, ohne einen vollständigen Phishing-Angriff simulieren zu müssen.

Die Heuristik kann für jede der möglichen Kollisionen genutzt werden. Die Wahl der genauen Kollision ist abhängig von dem Verfahren und welche Möglichkeiten es der Angreiferin bietet. Bei der Darstellung der E-Mail-Adresse ist es plausibel, dass $\mathcal{A}$ eine Nachricht mit einer selbstgewählten Darstellung der Herkunft konstruieren kann. Damit muss $\mathcal{A}$ im Kollisionsspiel SelVisCol gewinnen. Dies ist beispielsweise deutlich einfacher als auf eine Gleichheit bei zwei zufälligen Elemente zu hoffen.

E-Mail-Adressen

In vielen Anwendungen wird die Herkunft einer E-Mail nur über die Adresse dargestellt. Eine Angreiferin kennt diese Adresse und kann selbst eine beliebige, aber ungleiche E-Mail-Adresse wählen. Damit muss die Angreiferin eine ausgewählte visuelle Kollision finden.

Das bedeutet, dass die Angreiferin passend zu der Adresse von der legitimen Herkunft eine zweite Adresse konstruieren kann und einen starken Einfluss auf die Wahl hat. Die Angreiferin kann bei der Konstruk-

tion kognitive Eigenschaften ausnutzen. Durch die unterschiedlichen Kollisionsspiele wird deutlich, dass die Freiheit der Angreiferin bei der Darstellung der Herkunft einen Angriff deutlich vereinfacht. Dies ist eine allgemeine Gefahr eines Verfahrens.

Die Experimente zu Wahrnehmungsfehlern bei Texten, Zeichenketten und Buchstaben konnten immer einen Unterschied zwischen bestimmten Buchstaben aufzeigen.

Wort- und Buchstabensensitive Fehler

Die Erkennung einer Veränderung von Wörtern ist abhängig vom konkreten Wort und den beteiligten Buchstaben. Die Erkennung von Fehlern schwankt sehr.

Für die Darstellung der E-Mail-Adresse hat das Konsequenzen. Die Wahl der E-Mail-Adresse durch die legitime Partei, welche das Ziel eines Phishing-Angriffs ist, kann die Sicherheit gegen einen Phishing-Angriff verändern. Bei einer allgemeinen Sicherheitsanalyse von einem Verfahren ist dies eine besondere Herausforderung und erfordert eine Abschätzung der Sicherheit.

Es wird deutlich, dass die Darstellung einer E-Mail-Adresse fehleranfällig ist und zu gefährlichen Fehlern bei der menschlichen Wahrnehmung verleiten kann. Eine Angreiferin kann diese Fehler ausnutzen. Im folgenden Abschnitt werden noch weitere mögliche Fallstricke skizziert.

5.2.3 Weitere Fallstricke

Die Substitution von Buchstaben ist nicht der einzige Fallstrick bei der Wahrnehmung von E-Mail Adressen. Aus der Kognitionswissenschaft sind einige Effekte, welche das Lesen von Texten beeinflussen, bekannt. In diesem Abschnitt wird eine Auswahl von weiteren kognitiven Fallstricken betrachtet, um die Vielfältigkeit hervorzuheben. Diese Sammlung erhebt keinen Anspruch auf Vollständigkeit, aber verdeutlicht die umfangreichen Möglichkeiten.

Buchstabensalat

Wentura und Frings zeigen, dass Menschen Texte immer noch verständlich lesen können, obwohl die Buchstaben innerhalb der Worte vertauscht wurden, wobei der Anfangs- und Endbuchstabe unverändert blieb [107]. Dieser Effekt wurde in einer Studie von Rayner et al. genauer untersucht.

Kognitives Beispiel

Afugrnud enier Sduite an enier Elingshcen Unvirestiät ist es eagl, in wlehcer Rienhnelfoge die Bcuhtsbaen in eniem Wrot sethen, das enizg wcihitge dbaei ist, dsas der estre und lzete Bcuhtsbae am rcihgiten Paltz snid. Der Rset knan ttolaer Bölsdinn sien, und du knasnt es torztedm onhe Porbelme lseen. Das ghet dseahlb, wiel wir nchit Bcuhtsbae für Bcuhtsbae enizlen lseen, snodren Wröetr als Gnaezs. (Übernommen von Wentura und Frings [144].)

Mögliche Auswirkung auf eine Darstellung

Bei der Darstellung einer E-Mail-Adresse kann Buchstaben innerhalb der Domain verändert werden und so die Validierung der Domain verhindern.

Phishing Beispiel

Die Adresse `service@papyal.com` kann leicht mit `service@paypal.com` verwechselt werden.

Eine weitere Möglichkeit der Veränderung bei wenigen Buchstaben in einer Adresse ist die Möglichkeit, entweder einen Buchstaben hinzuzufügen oder zu streichen.

Fehlender Buchstabe

Menschen fallen Buchstaben in häufiger vorkommenden Wörtern seltener auf als in weniger häufigen, aber inhaltlich bedeutsamen Wörtern. Dieser Effekt wurde häufig untersucht, indem Studienteilnehmende bestimmte Buchstaben (z.B. *e* oder *t*) in einem Text markieren sollten.

Kognitives Beispiel

In der Studie von Corcoran wurde der Buchstabe *e* im Wort *the* am häufigsten im Vergleich zu anderen Wörtern mit einem *e* in einem Text nicht markiert [33]. Ähnliches hat Healy in ihrer Studie bei der Markierung vom *t* beobachtet [56].

Mögliche Auswirkung auf eine Darstellung

Adressen bestehen nicht nur aus Wörtern, sondern die *Top-Level-Domain* kommt sehr häufig vor und wirkt teilweise kryptisch. Eine Veränderung bei der *Top-Level-Domain* kann so leichter übersehen werden. Enthält eine Domain ein Füllwort oder Stoppwort, wie zum Beispiel *the, for* oder *of*, können Änderungen hier leichter übersehen werden.

Phishing Beispiel

Die Adresse `service@paypal.com` kann leichter mit `service@paypal.cm` als mit `service@paypa.com` verwechselt werden.
Die Adresse `service@bankofamerica.com` kann leichter mit `service@bankoamerica.com` als mit `service@banofamerica.com` verwechselt werden.

Bei diesem Effekt wird bereits deutlich, dass dieser in einem Angriff mit der Buchstabensubstitution kombiniert werden kann. Bei der Adresse `service@bankofamerica.com` kann beispielsweise das Wort *of* anstatt einem Substantive verändert werden und so ist eine mögliche Angriffsadresse `service@bankotamerica.com`.

5.3 Unterschiedliche Interpretationen

Eine Darstellung kann einer bestimmten Semantik folgen. Das `FROM`-Feld kann in den Anzeigenamen und die E-Mail-Adresse in den lokalen Teil, der *Top-Level-Domain*, *Second-Level-Domain* und den *Subdomains* unterteilt werden. Durch diese Interpretation bekommt die Darstellung eine Bedeutung. Für die Herkunft ist meistens die *Top-Level-Domain* sowie die *Second-Level-Domain* besonders wichtig. Eine fehlerhafte Interpretation der E-Mail-Adresse stellt damit eine besondere Gefahrenquelle dar. Wenn zwei E-Mail-Adressen dieselbe *Second-Level-Domain* aber unterschiedliche *Top-Level-Domains* haben, sollten diese nicht sofort der gleichen Herkunft zugeordnet werden. Beispielsweise sollte die E-Mail-Adresse `service@paypal.cn` nicht der Herkunft von `service@paypal.com` zugeordnet werden.

Die Interpretation einer Darstellung wird von den vorher beschrieben Prozessen beeinflusst und beeinflusst diese wiederum. Ein wichtiger Aspekt ist dabei das Wissen von $\mathcal{H}$ über die Semantik der Darstellung.

Eine Darstellung kann auf der Semantik aufgebaut sein und diese kann von $\mathcal{H}$ interpretiert werden. Das ist ein gefährlicher Angriffspunkt, weil ein Unterschied zwischen der tatsächlichen protokoll-basierten Semantik und der Interpretation vom Menschen $\mathcal{H}$ existieren kann und diese Auswirkungen auf die Sicherheit haben kann. Weitere Experimente simulieren dies. Die Anfrage meaning an $\mathcal{H}$ symbolisiert die Interpretation der Darstellung.

$\underline{\text{FixSemantic}^{\mathcal{A},\mathcal{H}^{\text{meaning}}}_{\text{Setup},\mathcal{L},s_1}}$

1 : $s_0 \leftarrow\!\!\$\ \mathcal{A}(s_1, \mathcal{L} \setminus s_1)$
2 : $b \leftarrow\!\!\$\ \{0,1\}$
3 : $\hat{b} \leftarrow \mathcal{H}^{\text{meaning}}(s_1, s_b)$
4 : **return** $\hat{b}, b$

$\underline{\text{Semantic}^{\mathcal{H}^{\text{meaning}}}_{\text{Setup},\mathcal{L}}}$

1 : $s_1 \leftarrow\!\!\$\ \mathcal{L}$
2 : $s_0 \leftarrow\!\!\$\ \mathcal{L} \setminus \{s_1\}$
3 : $b \leftarrow\!\!\$\ \{0,1\}$
4 : $\hat{b} \leftarrow \mathcal{H}^{\text{meaning}}(s_1, s_b)$
5 : **return** $\hat{b}, b$

Diese Konstruktionen folgen den vorherigen visuellen Kollisionen und $\mathcal{A}$ gewinnt genau dann, wenn $\hat{b} = 1$ und $b = 0$ die Rückgabewerte des Experiments sind. Wenn solche Kollisionen vorliegen, kann dies für einen Angriff auf das Verfahren angewendet werden und die Begründung dazu folgt der vorherigen Heuristik.

Unsicherheitsheuristik: Fehlerhafte Interpretation

Sei $\Pi = (\mathrm{R}_S, Gen)$ ein Darstellungsverfahren und $\mathcal{L}$ der Darstellungsraum von R_S und s_1 die Herkunftsdarstellung einer legitimen Nachricht m_1 von Setup. Dann ist Π unsicher, wenn eine Angreiferin $\mathcal{A}$ Folgendes kann:

1. Es ist plausibel, dass $\mathcal{A}$ das Experiment $\text{FixSemantic}^{\mathcal{A},\mathcal{H}^{\text{meaning}}}_{\text{Setup},\mathcal{L},s_1}$ gewinnt und sei s_0 die entsprechende Darstellung.
2. Es ist plausibel, dass $\mathcal{A}$ eine Nachricht mit der Herkunftsdarstellung s_0 erzeugen kann.

Im Kontext der E-Mail kann dies beispielsweise bedeuten, dass ein Wort vom `FROM`-Feld einer Nutzerin als Herkunft bekannt ist und die E-Mail dieser Herkunft zugeordnet wird [113]. Ein Beispiel dafür ist `norply@service-paypal.com` und in dieser E-Mail-Adresse wird das Wort PayPal wieder erkannt und darum der Herkunft PayPal zugeordnet. Erstaunlicherweise liefern unterschiedliche Implementierungen von URL-Parser unterschiedliche Ergebnisse [112]. Zusätzlich ist die Auswertung von manchen URLs nicht eindeutig. Beispielsweise kann `https://n.pr[0x00]@e.gg` zu `n.pr` oder `e.gg` ausgewertet werden, wobei eigentlich ein Fehler bei der URL auftreten sollte [112]. Die Mehrheit der von Reynolds et al. betrachteten Parser tut dies aber nicht [112]. Die Interpretation und das Parsen von URLs und von E-Mail-Adressen ist somit sowohl für den Menschen als auch die Maschine komplex und fehleranfällig [112].

5.4 Feldbeobachtungen

In den vorherigen Abschnitten wurden unterschiedliche Angriffsvektoren von Darstellungen mittels Zeichenketten illustriert und die Substitution von Buchstaben wurde als menschliche Fehlerquelle vorgestellt. In diesem Abschnitt werden tatsächliche vergangene Angriffe untersucht und Substitutionsangriffe genauer betrachtet. Die allgemeine Hypothese ist, dass kognitive Substitutionen häufiger als andere Substitutionen in Angriffsnachrichten vorkommen. Damit wird gezeigt, dass bei Angriffen die kognitiven Eigenschaften von Menschen ausgenutzt werden.

Bevor diese Forschungsfrage beantwortet werden kann, müssen Substitutionsangriffe aus vergangenen Angriffen extrahiert und analysiert werden.

Anwendungen können unterschiedliche Teile vom `FROM`-Feld darstellen und damit die Entscheidungsgrundlage für den Menschen sein. Für einen Substitutionsangriff ist insbesondere die *Second-Level-Domain* relevant, weil Sicherheitsmechanismen hier eine identische Täuschung verhindern können und eine Wahrnehmungstäuschung eine Alternative sein kann. In gängigen Anwendungen wird dieser Teil vom `FROM`-Feld angezeigt und ist besonders bedeutsam zur Bestimmung der Herkunft einer Nachricht. Aus diesem Grund wird die nachfolgende Analyse von Substitutionsangriffen auf die *Second-Level-Domain* beschränkt. Natürlich sind diese Angriffe auf andere Bereiche vom `FROM`-Feld anwendbar.

Es ist unklar, welcher Teil vom `FROM`-Feld dem Menschen tatsächlich angezeigt wird, aber zur Vereinfachung der Analyse wurde nur die *Second-Level-Domain* betrachtet. Bei der Betrachtung von den weiteren Feldern wird der Bezug zu den einzelnen E-Mails schwerer nachvollziehbar. Das Programm zur späteren Analyse wurde bereits auf die Analyse aller Teile vom `FROM`-Feld angepasst, aber es fehlt der Bezug zu den jeweiligen E-Mails.

5.4.1 Methode

Zur Erkennung von Substitutionen muss die legitime Herkunft mit der entsprechenden Domain bekannt sein. Für die Herkunft bzw. die entsprechende Domain einer unbekannten oder bösartigen E-Mail kann untersucht werden, ob eine Substitution mit einer legitimen Herkunft vorhanden ist. Zur Untersuchung von praktischen Angriffen im Kontext von E-Mails muss berücksichtigt werden, dass beim `FROM`-Feld alle Teile aus mehreren Wörtern bestehen können. Mehrere Wörter innerhalb einer Domain werden üblicherweise durch Bindestriche (−) getrennt, da Leerzeichen nicht verwendet werden können. Für die nachfolgende Analyse hat ein Substitutionsangriff folgende Eigenschaften:

1. Ein Wort aus der *Second-Level-Domain* unterscheidet sich von einer legitimen Herkunft um genau ein Zeichen.
2. Der Unterschied zwischen den beiden Wörtern ist nur der Austausch von einem Zeichen mit einem anderen Zeichen.

Eine Substitution wird als kognitive Substitution bezeichnet, wenn die Substitution des Buchstaben mit dem anderen in der Literatur beschrieben wird. Als Literaturgrundlage werden Haley [57] sowie Bole und Clifford [19] berücksichtigt.

Die Erkennung von Substitutionsangriffen erfolgt wie folgt:

Algorithm 1 Substitutionserkennung

```
Input: F: List of email-adresses, T: List of target-strings
subs ← {}
for all from∈ F do
  d_0, ... d_{n-1} ← from.split('.')
  // Consider co.uk etc.
  if |d_{n-2}| ≤ 3 then
    snd ← d_{n-3}
  else
    snd ← d_{n-2}
  end if
  words ← snd.split('-')
  for all w ∈ words do
    for all t ∈ T do
      dist ← editdistance(w,t)
      if e = 1 ∧ |t| = |w| then
        subs.append(from)
      end if
    end for
  end for
end for
return subs
```

Im Anschluss von Algorithmus 1 wird untersucht, ob diese Substitution aus der Kognitionswissenschaft bekannt ist oder nicht.

Der wesentliche Teil der Domain zur Bestimmung der Herkunft ist die *Second-Level-Domain*, weil oftmals die Registrierung einer Domain unter einer *Top-Level-Domain* erfolgt. Allerdings gibt es Sonderfälle, wie zum Beispiel `co.uk`. In diesen Fällen leitet sich die Herkunft nach der ersten *Subdomain* ab. In wenigen Fällen erfolgt eine Registrierung einer eigenen Domain als *Subdomain* unter der *Second-Level-Domain*. Die Bestimmung der Geltungsbereiche ist nicht immer eindeutig möglich. Ein gängiges Verfahren dazu ist der Abgleich mit einer tagesaktuellen Webseite[2]. Diese wird von Google, Mozilla und anderen dazu verwendet und regelmäßig aktualisiert. Eine historische Betrachtung ist mit dieser Liste

[2] www.publicsuffix.org/

nicht möglich, weil sich diese ändert. Zur Vereinfachung werden stattdessen der letzte und, falls dieser sehr kurz (weniger oder gleich drei Zeichen) ist, der vorletzte Teil einer Domain als *Top-Level-Domain* interpretiert. Aus dem `FROM`-Feld wurden zuvor die E-Mail-Adresse und der Anzeigename extrahiert und unnötige Leerzeichen sowie Trennzeichen (z.B. $<, >$) wurden entfernt.

Die Menge legitimer Ziele (T) muss vorher manuell erstellt werden. In der Praxis kann diese aus der Historie abgeleitet und leicht angepasst werden. Dem Datensatz sind keine Angriffsziele beigelegt. Aus diesem Grund müssen die möglichen Ziele mit einer legitimen Herkunft aus dem Datensatz ausgewählt werden. Dafür wurde der Datensatz zusammengefasst und manuell gesichtet. Zur manuellen Auswahl wurde nach unterschiedlichen Zeichenketten im `FROM`-Feld gesucht und innerhalb der Domain nach einem Bereich mit einer Textlänge von 4 oder länger gesucht. Diese Domainteile wurden mittels Editierabstand (auch bekannt als Levenshtein Distanz [82]) verglichen und wenn der Abstand kleiner gleich drei war zusammen gruppiert. In den Gruppen werden manuell bekannte Organisationen identifiziert.

Im nächsten Schritt werden dann mit dem obigen Algorithmus die `FROM`-Felder untersucht. Der Algorithmus wurde mittels Python implementiert und ist öffentlich verfügbar[3].

5.4.2 Durchführung

Das Cambridge Cybercrime Centre[4] sammelt und erhebt unterschiedliche Daten in Bezug auf Online-Kriminalität. Unter anderem wird eine Sammlung von Phishing-Nachrichten seit 2005 zur Verfügung gestellt. Dieser Datensatz umfasst E-Mails mit allen Meta-Daten. Dies ermöglicht insbesondere die Betrachtung der `FROM`-Felder. Diese E-Mails wurden mittels eines *Honeypot* gesammelt und geben somit keine Anhaltspunkte über die Erfolge der jeweiligen Angriffe. Die E-Mails wurden manuell in die Kategorien legitime E-Mails, E-Mails mit Malware, Phishing-E-Mails, Scams und Spam eingeteilt. Der genutzte Datensatz umfasst nur die Phishing-Angriffe

Dem Datensatz sind nicht die legitimen Absender beigefügt. Diese sind jedoch zur Analyse nötig. Nach einer manuellen Sichtung wurden 47 mögliche legitime Absender von Nachrichten erkannt. Diese sind populäre Vertreter aus der Finanz- und IT-Industrie, wie zum Beispiel Paypal, Amazon, Ebay, Citibank und Apple.

Einige `FROM`-Felder beinhalten das Wort *email*, welches eine Substitution mit *gmail* darstellt. Bei der manuellen Durchsicht wurde aber festgestellt, dass diese Zuordnung häufig falsch ist. Aus diesem Grund wurden Substitutionen mit dem Wort *mail* nicht berücksichtigt. Das Wort *cahoo* wurde mehrfach gefunden und hat einen Editierabstand von Eins zu den Wörtern *cahoot* sowie *yahoo*. Aus diesem Grund wurde dieses Wort für die Analyse nicht berücksichtigt. Ebenso wurden zwei Substitutionen mit einem Nicht-ASCII Zeichen identifiziert. Das sind homoglyphische Angriffe mit unterschiedlichen Alphabeten; diese wurden ebenso ignoriert.

[3] https://git.imp.fu-berlin.de/wieseoli/sender_security.git
[4] https://www.cambridgecybercrime.uk/

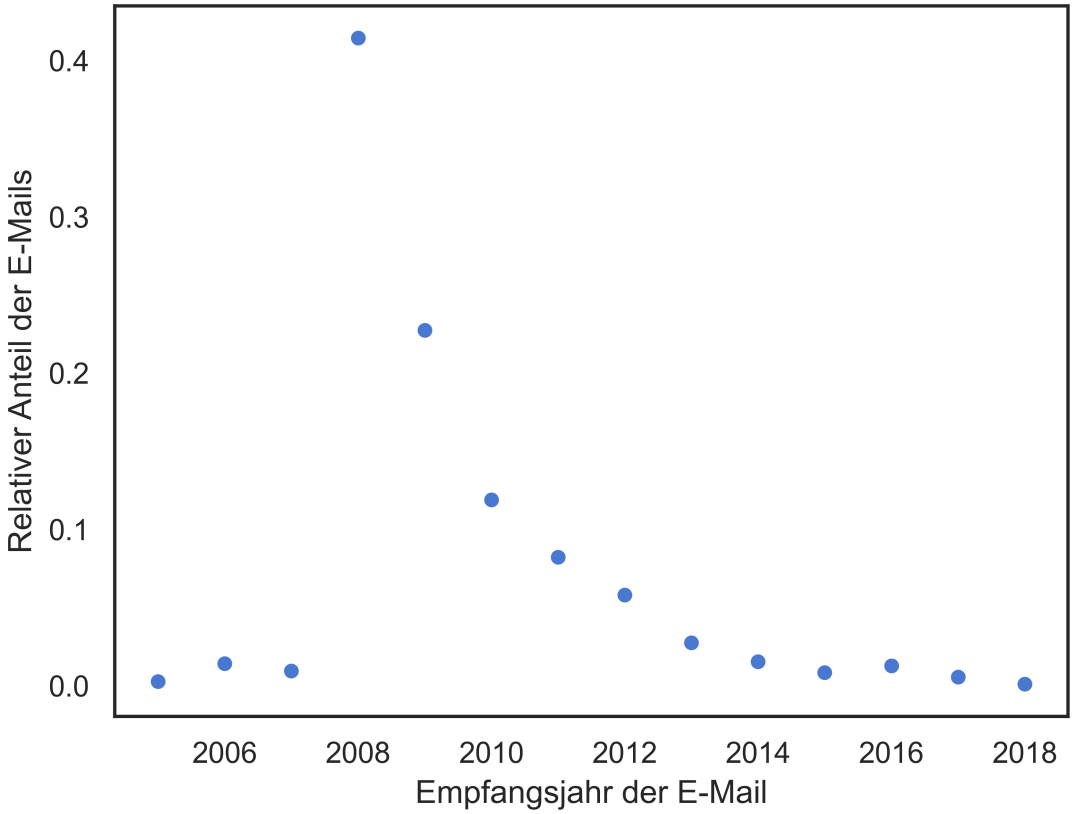

Abbildung 5.2: Die Verteilung der E-Mails im CCC-Datensatz nach Jahren. Ein großer Teil der E-Mails stammt aus den Jahren 2009 bis 2012.

5.4.3 Ergebnis

Der Datensatz umfasst 63.449 E-Mails mit insgesamt 24.031 unterschiedlichen `FROM`-Feldern. Die Abbildung 5.2 zeigt die Verteilung der unterschiedlichen `FROM`-Felder pro Jahr. Durchschnittlich wurden $2,64$ E-Mails pro `FROM`-Feld (std $= 19,95$) empfangen.

Aus der *Second-Level-Domain* wurden 3.774 unterschiedliche Wörter extrahiert und jeweils der Editierabstand mit den Zielen bestimmt. 143 Wörter waren exakt die gleichen zu einem Ziel (Editierabstand Null). 126 Wörter hatten einen minimalen Editierabstand von Eins zu einem Ziel. Davon wurde bei 40 Wörtern ein Buchstabe hinzugefügt und bei 40 Wörtern wurde ein Zeichen entfernt. Bei 46 wurde genau ein Zeichen ausgetauscht. Der minimale durchschnittliche Editierabstand zu einem Ziel beträgt $5,1$ bei einer Standardabweichung von 3. Abbildung 5.3 stellt die Verteilung des minimalen Editierabstands dar.

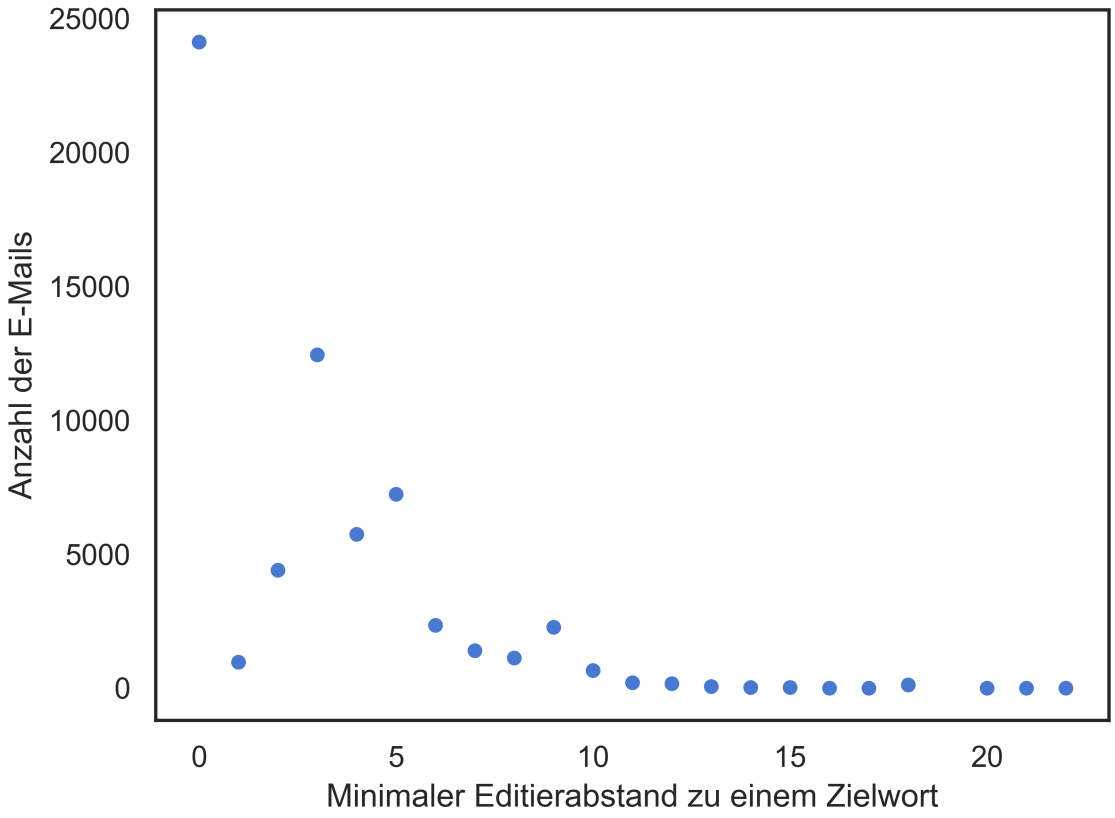

Abbildung 5.3: Der minimale Editierabstand aus einem Wort der E-Mail-Adresse zu einer der möglichen Zielnamen. Ein großer Teil der Adressen aus dem `FROM`-Feld einer E-Mail beinhaltet bereits das eigentliche Ziel.

Ein Drittel aller *Second-Level-Domain* beinhaltet ein Wort mit einem minimalen Editierabstand von 0 zu einem Ziel.

Angriffswörter

Es wurden 32 unterschiedliche Angriffswörter für einen Substitutionsangriff auf der Ebene der *Second-Level-Domain* identifiziert. Tabelle 5.1 listet alle gefundenen Angriffswörter auf. 18 Wörter sind einer bekannten Substitution aus der Kognitionswissenschaft zugeordnet und 14 Wörter sind unbekannte Buchstabenvertauschungen. Jedes dieser Angriffswörter kann in mehreren E-Mails vorkommen. Durchschnittlich kommt eine kognitive Substitution in $13,5$ E-Mails vor (std: $26,42$). Anderen Substitutionen werden durchschnittlich nur $1,92$ E-Mails zugeordnet (std: $1,24$).

Zeichenkette	Ziel	Bekannte kog. Substitution?	# Absenderin	# E-Mails
mbay	ebay	Nein	1	1
postbanc	postbank	Nein	1	1
dll	dhl	Nein	1	2
barklays	barclays	Nein	1	1
ehay	ebay	Nein	2	2
e8ay	ebay	Nein	1	1
app1e	apple	Nein	1	1
paypol	paypal	Nein	1	1
ebuy	ebay	Nein	3	5
halifix	halifax	Nein	1	1
paypak	paypal	Nein	1	1
palpal	paypal	Nein	2	2
eboy	ebay	Nein	1	2
paypa1	paypal	Nein	2	2
haiifax	halifax	Ja	2	6
paypai	paypal	Ja	25	62
biockchain	blockchain	Ja	1	2
appte	apple	Ja	1	1
barciays	barclays	Ja	4	6
appie	apple	Ja	1	1
maslercard	mastercard	Ja	1	1
amezon	amazon	Ja	2	2
ebav	ebay	Ja	1	1
eday	ebay	Ja	1	1
payppl	paypal	Ja	1	1
santender	santander	Ja	1	1
peypal	paypal	Ja	2	2
paypel	paypal	Ja	2	2
netwest	natwest	Ja	1	1
ebey	ebay	Ja	10	11
bankotamerica	bankofamerica	Ja	1	1
pavpal	paypal	Ja	1	6

Tabelle 5.1: Es werden die gefundenen Substitutionen darstellt. Die relevante (Teil)-Zeichenkette wurde in einer Adresse im Datensatz erkannt. Das Ziel ist die Zeichenkette, welche sich um ein Zeichen mit der (Teil)-Zeichenkette unterscheidet. Die Anzahl der unterschiedlichen Absenderinnen und die Anzahl der E-Mails verdeutlicht die Häufigkeit der gefundenen Substitution.

Durchschnittlich war ein Wort bei den Substitutionsangriffen $6,16$ Zeichen lang (std: $2,19$). Bei einem Angriffswort erfolgte die Substitution an der ersten Stelle und bei 5 an der letzten Stelle vom Wort. In Relation zum Wort erfolgte die Substitution durchschnittlich in der Mitte vom Wort (durchschnittlich: $0,47$ und std: $0,24$)

Abbildung 5.4 stellt die relative Position der Substitutionen in Bezug auf die Häufigkeit dar. Es ist auffällig, dass die meisten Substitutionen tendenziell in der zweiten Hälfte der Wörter erfolgten.

Buchstabensubstitutionen

Die Angriffe können nicht nur auf der Ebene der Wörter betrachtet werden, sondern auf der Ebene der einzelnen Buchstabensubstitutionen. Tabelle 5.2 listet alle gefundenen Buchstabensubstitutionen auf. 8 Buchstabenvertauschungen sind aus der Literatur der Kognitionswissenschaft bekannt und 12 sind unbekannt. Es gab kognitive Buchstabenvertauschungen, welche gegen 5 verschiedene Ziele eingesetzt wurden.

Insgesamt 131 diese Substitutionsangriffe wurden in den E-Mails identifiziert. Davon sind 108 Angriffe einer Substitution aus den Kognitionswissenschaft bekannt und 23 sind unbekannt.

Kognitive Substitutionen wurden vereinzelt gegen 5 unterschiedliche Ziele, also legitime Organisationen, verwendet. Bei den unbekannten Substitutionen waren es hingegen maximal nur 2 unterschiedliche Ziele je Substitution. Ein ähnliches Verhalten zeigt sich bei der Häufigkeit in E-Mails. Durchschnittlich kam eine kognitive Substitution in $13,5$ E-Mails vor (std: $26,42$) und die anderen Substitutionen in durchschnittlich $1,92$ E-Mails (std: $1,24$).

Ziele

Insgesamt konnten gegen 13 unterschiedliche Ziele Substitutionsangriffe beobachtet werden. Tabelle 5.3 ordnet die unterschiedlichen Angriffe den Zielen zu.

Durchschnittlich erhielt ein Ziel $8,31$ E-Mails mit einer bekannten Substitution in der *Second-Level-Domain* (std: $19,77$), aber nur $1,77$ E-Mails mit einer anderen Substitution (std: $3,22$). Bei der Betrachtung des relativen Anteils an kognitiven Substitutionen je Ziel waren durchschnittlich $0,76$ eine kognitive Substitution (std: $0,37$). Der Mittelwert über alle Ziele ist in Tabelle 5.3 dargestellt.

5.4.4 Diskussion

Der Datensatz besteht im Großteil aus E-Mails aus den Jahren 2008/2009 (vgl. Abbildung 5.2) und im Vergleich zur technologischen Schnelllebigkeit und die fortschreitende Verbreitung von Sicherheitsprotokollen (unter anderem SPF und DKIM) ist dieser Datensatz eher von historischer Bedeutung. Es sollte berücksichtigt werden, dass aktuelle Angriffe somit anders aussehen können. Für die konkrete

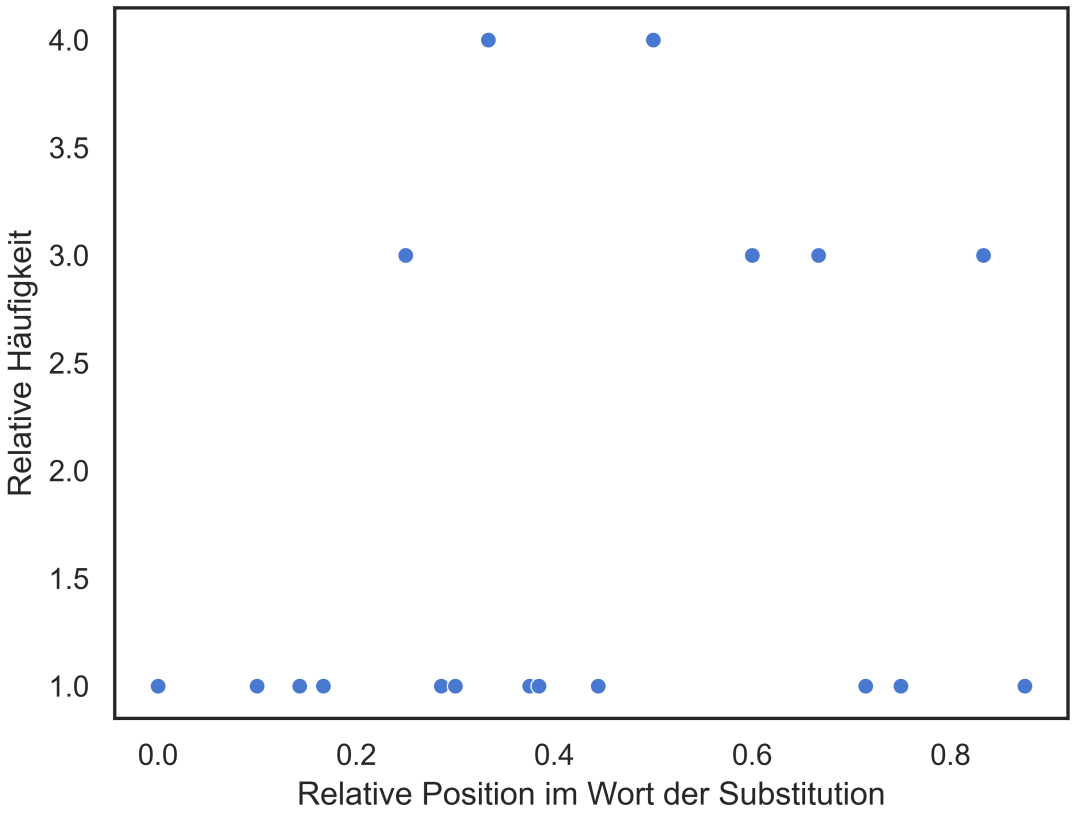

Abbildung 5.4: Es wird die relative Position innerhalb des Wortes der Substitution und deren Häufigkeit dargestellt. Besonders häufig erfolgt eine Substitution im mittleren Bereiches des Wortes, also zwischen 0,3 und 0,6. Eine Substitution genau in der Mitte ist nicht immer möglich, weil nicht alle Buchstaben gleich gut für eine Substitution geeignet sind.

Substitution		kognitive	Anzahl der verschiedenen		
von	zu	Substitution?	Ziele	FROM-Felder	E-Mails
a	e	Ja	5	18	19
a	i	Nein	1	1	1
a	o	Nein	2	2	3
a	p	Ja	1	1	1
a	u	Nein	1	3	5
b	8	Nein	1	1	1
b	d	Ja	1	1	1
b	h	Nein	1	2	2
c	k	Nein	1	1	1
e	m	Nein	1	1	1
f	t	Ja	1	1	1
h	l	Nein	1	1	2
k	c	Nein	1	1	1
l	l	Nein	2	3	3
l	i	Ja	5	33	77
l	k	Nein	1	1	1
l	t	Ja	1	1	1
t	l	Ja	1	1	1
y	l	Nein	1	2	2
y	v	Ja	2	2	7

Tabelle 5.2: Es werden die verschiedenen gefundenen Substitutionen und deren Vorkommen dargestellt. Beispielsweise wurde in der ersten Substitution der Buchstabe a durch den Buchstaben e ausgetauscht. Dies ist eine bekannte kognitive Substitution und wurde insgesamt in 19 E-Mails bei 5 unterschiedlichen Zielen, also zum Beispiel PayPal, durchgeführt.

Ziele	Unterschiedliche FROM-Felder				Unterschiedliche E-Mails			
	kog.	¬ kog.	Gesamt	Anteil	kog.	¬ kog.	Gesamt	Anteil
amazon	2	0	2	1,00	2	0	2	1,00
apple	2	1	3	0,67	2	1	3	0,67
bankofamerica	1	0	1	1,00	1	0	1	1,00
barclays	4	1	5	0,80	6	1	7	0,86
blockchain	1	0	1	1,00	2	0	2	1,00
ebay	12	8	20	0,60	13	11	24	0,54
halifax	2	1	3	0,67	6	1	7	0,86
mastercard	1	0	1	1,00	1	0	1	1,00
natwest	1	0	1	1,00	1	0	1	1,00
paypal	31	6	37	0,84	73	6	79	0,92
santander	1	0	1	1,00	1	0	1	1,00
dhl	0	1	1	0,00	0	2	2	0,00
postbank	0	1	1	0,00	0	1	1	0,00
Total	58	19	77	0,75	108	23	131	0,82

Tabelle 5.3: Für die gefundenen Ziele wird jeweils die Häufigkeit unterschiedlicher Substitutionen dargestellt. Es wird dabei nach der Anzahl der E-Mails und der Anzahl an verschiedenen FROM-Feldern unterschieden. Die Anzahl der kognitiven Substitutionen sind in den Spalten *kog.* und die Anzahl der nicht-kognitiven in der Spalte ¬ *kog.*. Der Anteil ist der Anteil von kognitiven Substitutionen. In fast allen Zielen gab es mehr kognitive Substitutionen als nicht-kognitive Substitutionen.

Forschungsfrage ist dies aber ausreichend, weil der historische Nachweis der Nutzung von kognitiven Schwächen bereits sinnvoll und ausreichend ist.

Der Datensatz umfasst zwar 63.449 E-Mails, aber dennoch konnten nur 32 Angriffswörter mit einem Editierabstand von Eins identifiziert werden. Mehr als ein Drittel aller E-Mails hat bereits die Zieladresse als Wort in der *Second-Level-Domain* und also einen minimalen Editierabstand von Null. In diesen E-Mails wurde entweder die Zieladresse mit einem Bindestrich um weitere Wörter ergänzt oder die Zieladresse ist ohne Veränderungen übernommen werden.

Die Anzahl der E-Mails mit einem Editierabstand von zwei bis neun sind jeweils häufiger vertreten als mit einem Editierabstand von Eins. Eine Ursache kann dafür sein, dass die Täuschung nicht nur durch die *Second-Level-Domain* erfolgte, sondern durch weitere Teile vom `FROM`-Feld und insbesondere vom Anzeigenamen erfolgte.

Es gibt nur eine Zeichenkette mit einer möglichen Kollision zwischen möglichen Zielen bei einem Editierabstand von Eins. Das betrifft den erwähnten Konflikt zwischen Yahoo und Cahoot. Dieser Fall wurden ebenso wie die Konflikte mit Gmail bei der Analyse ignoriert.

Ziele

Nicht zu allen 47 möglichen Zielen konnten Substitutionsangriffe gefunden werden und die Varianten und die Häufigkeit der Angriffe sind sehr unterschiedlich. In der Tendenz basierten die Angriffe mehrheitlich auf kognitiven Substitutionen anstatt auf anderen Substitutionen. Insbesondere Ebay und PayPal waren deutlich stärker von Substitutionen betroffen. Eine Erklärung ist dafür, dass beide Unternehmen ein häufiges Ziel von Phishing-Angriffen waren und oftmals eine Vorreiterrolle bei der Bekämpfung eingenommen haben. Beispielsweise hat Gmail 2008 angekündigt, fragwürdige E-Mails von beiden Domains nicht mehr anzuzeigen.[5] In der Folge wurden wahrscheinlich andere Angriffstechniken genutzt und insbesondere Substitutionen sind eine Alternative zu einer direkten Fälschung. Die Hälfte aller betrachteten Substitutionsangriffe zielten auf PayPal ab und mit *paypel, peypal, paypai, payppl und pavpal* wurden einige mögliche Varianten gefunden. Dies sind nur 5 von insgesamt 31 möglichen kognitiven Substitutionen von PayPal. Es wurden somit nicht alle möglichen Varianten ausgewählt. Insbesondere der Anfangsbuchstabe wurde nicht verändert. Im Gegensatz dazu wurden mit *paypa1, paypak, palpal, paypol, paypay* andere Substitutionen gewählt. Diese Substitutionen sind deutlich seltener im Datensatz zu finden.

Substitutionen im Detail

Es gibt Substitutionen von einem Buchstaben durch eine Ziffer. Im Datensatz sind *e8ay* statt ebay und *app1e* statt *apple* vorhanden. Dies stellt eine Einschränkung zu den betrachteten kognitiven Substitutionen dar, denn diese beruhen auf Zeichen vom selben Alphabet. Andere nicht-kognitive Substitutionen

[5] https://gmail.googleblog.com/2008/07/fighting-phishing-with-ebay-and-paypal.html

(*eboy, ebuy, eday* statt ebay, *netwest* statt natwest, *halifix* statt halifax) bestehen aus neuen Namen und können andere kognitive Effekte, welche auf Wörtern statt auf zufälligen Zeichenketten basieren, nutzen. Eine alternative Erklärung ist die Verwirrung der Nutzerinnen bei der Interpretation vom Wort, also der Adresse.

Auffällig ist, dass die Substitutionen häufig in der zweiten Hälfte vom Wort stattfanden. Eine mögliche Erklärung ist, dass eine Substitution in der zweiten Worthälfte erfolgsversprechender ist und seltener wahrgenommen wird. Dies kann dadurch erklärt werden, dass Menschen nur eine bestimmte Anzahl an Zeichen auf einmal wahrnehmen können. Dies wird von Rayner als die Theorie vom *Moving-Window* beschrieben [105]. Mit der Erwartungshaltung und Vorwissen über die mögliche Herkunft ist es damit denkbar, dass die Fehler seltener wahrgenommen werden und Menschen eher getäuscht werden.

Insgesamt wurden 18 von 87 bekannten kognitiven Substitutionen gefunden. Hierbei muss berücksichtigt werden, dass nicht alle kognitiven Substitutionen in den Zeichenketten angewendet werden können. Es wurden nur Substitutionen mit kleinen Buchstaben berücksichtigt. Insgesamt wurden 14 unterschiedliche nicht-kognitive Substitutionen gefunden. Die Anzahl der möglichen Substitutionen ist durch die Anzahl der möglichen Zeichen (also dem lateinischen Alphabet und Zahlen) beschränkt und kann somit abgeschätzt werden durch mindestens $(26 + 10) \cdot (26 + 10 - 1) = 1.260$. Bei nicht-kognitiven Substitutionen gibt es nur 3 mit zwei oder mehr versendeten E-Mails, aber bei den kognitiven Substitutionen trifft dies auf 7 unterschiedliche zu. Eine mögliche Erklärung ist, dass die kognitiven Substitutionen erfolgreicher waren und darum häufiger gewählt wurden oder die nicht-kognitiven wurden eher zufällig gewählt. Die kognitiven Substitutionen wurden hingegen von unterschiedlichen Personen wiederholt gewählt. Die Anzahl der unterschiedlichen Substitutionen unterscheidet sich nicht zwischen kognitiven und nicht-kognitiven Substitutionen, aber in der Häufigkeit gibt es zwischen diesen Gruppen Unterschiede. 108 Angriffe einer Substitution aus den Kognitionswissenschaft und 23 andere Substitutionsangriffe konnten identifiziert werden. Ein Binomialtest zum Vergleich der Häufigkeitsverteilung zeigt einen signifikanten Unterschied zur Gleichverteilung (p-Wert: $2 \cdot 10^{-14}$). Diese historischen Angriffe verdeutlichen, dass die Angriffe eine mögliche kognitive Schwäche ausnutzen. Vermutlich wurden zur Konstruktion der Angriffe keine wissenschaftlichen Arbeiten zur Kognition berücksichtigt, sondern die Substitutionen wurden adhoc und intuitiv gebildet. Die am meisten verwendeten Substitutionen sind naheliegend. Dennoch gibt es Parallelen zu den Erkenntnissen aus der Kognitionswissenschaft. Die kognitiven Eigenschaften (und Einschränkungen) von Menschen bei der Benutzung der Verfahren sollten bereits bei der Konstruktion und Implementierung berücksichtigt werden. Ansonsten werden dies die Angreiferinnen tun.

5.4.5 Einschränkungen

Die Analyse umfasst einen Ausschnitt von möglichen Angriffen und der Auswahlprozess ist nur eingeschränkt, denn die E-Mails stammen aus einer Sammlung einer einzelnen Person. Ein Großteil der E-Mails ist mehr als zehn Jahre alt und damit eher von historischer Bedeutung. Damals war die Verifikation mittels DKIM oder anderen kryptographischen Ansätzen unüblich. Dies spiegelt sich in den vielen

identischen Fälschungen in dem Datensatz wider. Die aktuelle Spam- und Phishingerkennung von aufmerksamen E-Mail-Providern verhindert einen Großteil dieser Angriffe[6] und damit sind keine Schlussfolgerungen auf die aktuelle Situation möglich. Die Analyse von anderen Angriffen ist unvollständig und bereits in der Diskussion wurden weitere mögliche kognitive Effekte diskutiert, welche die Erkennung einer Täuschung beeinflussen können. Zusätzlich wurden Angriffe basierend auf speziellen Encodings oder Alphabeten nicht betrachtet. Der Umgang mit diesen Angriffen ist sehr stark von der Darstellung einer E-Mail-Anwendung abhängig. Dies betrifft die Unterschiede in Bezug auf Schriftart und Schriftgröße, welche bei Veröffentlichungen aus der Kognitionswissenschaft immer als Teil des Versuchsaufbaus angegeben wurden.

5.5 Verwandte Arbeiten aus der Sicherheitsforschung

In der Sicherheitsforschung gibt es bereits eine Vielzahl von Veröffentlichungen zu gefälschten URLs. Vieles lässt sich dabei auf die E-Mail übertragen.

Neben der visuellen Ähnlichkeit ist ein weiteres Maß für Webseiten die Distanz zwischen zwei Buchstaben im Tastaturlayout. Wenn eine Person eine bestimmte URL besuchen möchte und diese eintippt, kann ein Buchstabe vergessen werden oder ein auf dem Tastaturlayout gelegener Nachbarbuchstabe eingetippt werden. Diese Angriffe werden als *Typosquatting*-Angriffe bezeichnet [2].

Reynolds et al. untersuchten die Wirkung von solchen und ähnlichen Angriffen, in denen Buchstaben vertauscht, weggelassen oder hinzugefügt werden, auf den Menschen [113]. In ihrer Studie ordnete ein Drittel der Teilnehmenden die URL `twittter.com` dem Unternehmen zu, obwohl die Domain ein *t* zu viel umfasst. Dieses Beispiel ist sehr ähnlich zu der Substitution von Buchstaben und verwendet ähnliche Muster.

Die Internationalisierung von Domains und die Zulassung von unterschiedlichen Alphabeten ermöglicht so genannte homoglyphische Angriffe [65, 73, 102, 113]. In diesen Angriffen werden Zeichen aus einem Alphabet durch sehr ähnlich aussehende Zeichen eines anderen Alphabets ersetzt. Abbildung 5.5 stellt ein Beispiel, in dem nur ein Zeichen ausgetauscht wurde, dar. In einer Studie von Reynolds et al. erkannte die Hälfte der Studienteilnehmenden in einer URL den homoglyphischen Angriff [113]. Die Unterschiede zwischen Buchstaben in verschiedenen Alphabeten schwankt sehr.

Tian et al. haben mittels OCR und anderen Techniken ein auf maschinelles Lernen aufbauendes System gebaut, um Phishing-Webseiten zu erkennen [132]. Sie haben insbesondere die Darstellung der URL berücksichtigt. Die Erkennung von Phishing-Webseiten mittels maschinellen Lernens wird nicht berücksichtigt, weil die Erkennung ohne die Entscheidung der Nutzerinnen erfolgt.

Althobaiti et al. zeigen die hohe Komplexität bei der Interpretation einer URL bzw. Domain und nennen einige mögliche Angriffe [6]. Sie untersuchen mögliche Hilfestellungen für Menschen zur Erkennung

[6] Dies wird beispielsweise an der rückläufigen Zahl an E-Mails pro Jahr deutlich.

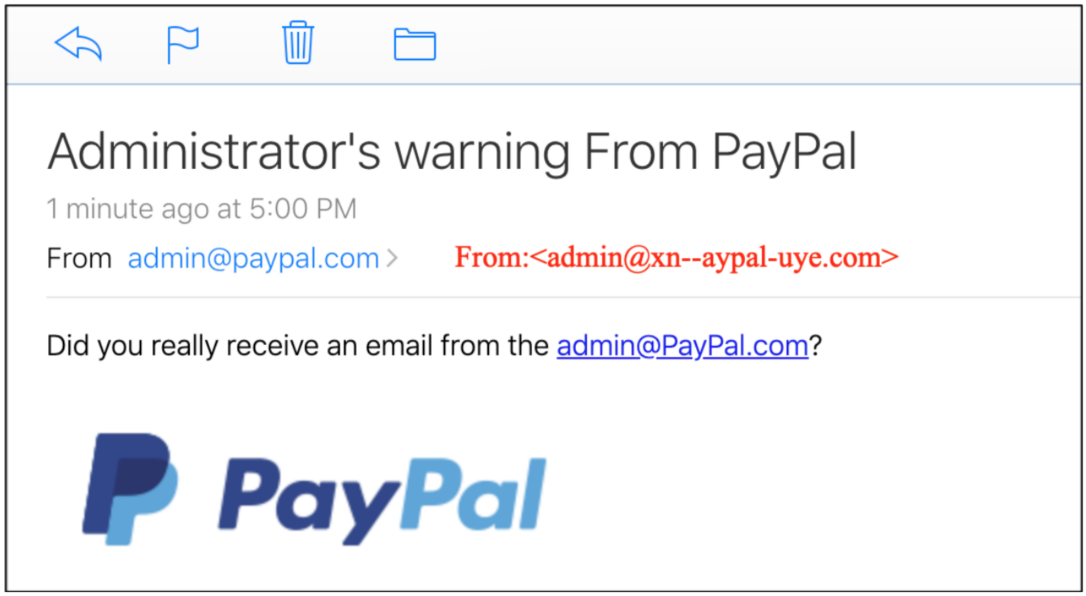

Abbildung 5.5: Ein homoglpyhischer Angriff wird in dem Web-Interface von iCloud dargestellt. Die Grafik wurde von Shen et al. übernommen [65]. Das rote `FROM`-Feld zeigt die tatsächliche E-Mail-Adresse an.

von Angriffen und folgern am Ende:

> Some participants still had difficulty understanding the more complex concepts such as PageRank and location.

Volkmer et al. haben mit Torpedo ein Tool konstruiert und untersucht, welches Menschen bei der Interpretation von URLs unterstützt [136]. Petelka et al. haben in ähnlicher Weise Warnungen bei URLs als Gegenmaßnahme zu Phishing-Angriffen untersucht [99]. Beide Ansätze unterstützen Menschen bei der Interpretation einer URL, haben aber den allgemeinen Nachteil einer Habituation bei Warnungen bei einer gleichzeitig hohen falschen Positivrate.

Neben menschlichen Schwächen bei der Betrachtung einer E-Mail wurden Angriffe auf die Protokolle und Spezifizierungen der E-Mail publiziert. Shen et al. haben die Gefahren von mehreren `FROM`-Feldern und ähnlichen Angriffen untersucht [65]. Hu und Wang haben die Gegenmaßnahmen von E-Mail-Servern und -Providern untersucht [62]. Dabei war weniger die Einrichtung von DKIM oder SPF das Problem, sondern vielmehr die Umsetzung der Richtlinien bei eingehenden E-Mails. E-Mail-Provider stehen beim automatischen Löschen von als Angriff erkannten E-Mails vor großen Herausforderungen, da fälschlicherweise gelöschte E-Mails negative Auswirkungen auf die Nutzungserfahrung haben und großes Frustrationspotential für Nutzerinnen haben. Gleichzeitig kennt der eingehende E-Mail-Provider nicht die gesamte Serverkonfiguration und die verwendeten Sicherheitsmaßnahmen vom ausgehenden

E-Mail-Provider.

5.6 Zusammenfassung

Eine Darstellung der Herkunft mittels Zeichenketten impliziert viele sehr unterschiedliche Gefahren und Tücken. Einerseits ist die Darstellung von Zeichen in der Praxis mit verschiedenen Kodierungen, Alphabeten, Schriftarten und spezifischen Eigenheiten, wie zum Beispiel Wechsel der Schriftrichtung, nicht trivial und bietet vielseitige Möglichkeiten den Menschen zu täuschen. Daneben gibt es eine Vielzahl an kognitiven Einschränkungen, wenn Menschen Zeichenketten wahrnehmen, lesen und verarbeiten. Insbesondere bei einer kurzen Betrachtungszeit der Zeichenketten ist ein erfolgreicher Angriff plausibel. In diesem Kapitel wurden einige kognitive Fallstricke für E-Mail-Adressen dargestellt und diese Fallstricke konnten in historischen Angriffen erkannt werden. Neben den erwähnten kognitiven Fallstricken gibt es weitere potentielle Gefahren, wenn Menschen Texte lesen. Die bestehende Literatur zum Lesen und Verarbeiten von Texten und Zeichenketten ist historisch gewachsen und immer noch ein aktuelles Forschungsfeld aus der Kognitionswissenschaft. Teilweise ist dieses Wissen in Lehrbüchern [58, 83, 103, 127, 141, 144] oder in Übersichtsartikeln [32, 104, 106] zusammengefasst.[7] Der vorgestellte Datensatz bietet sich an, um weitere Fallstricke zu bestätigen. Beispielsweise sind phonetisch ähnliche Wörter anfällig für Verwechslungen [127].

Im Gegensatz zu der bisherigen Forschung im Sicherheitskontext wurde der Bezug zur Kognitionswissenschaft und deren Erkenntnissen hervorgehoben. Es wird damit deutlich, dass bei der Sicherheitsanalyse eines Verfahrens Erkenntnisse aus der Kognitionswissenschaft berücksichtigt werden müssen und dieser Hinweise auf mögliche Schwächen eines Systems geben.

Die Algorithmen zur Analyse können für eine einfache Gegenmaßnahme adaptiert werden, aber die Effektivität dieser ist unklar und weitere Angriffe mittels anderer Schwächen der menschlichen Kognition sind möglich. Die Erkenntnisse aus diesem Kapitel sind auf die Darstellung von URLs übertragbar und lassen sich somit auf den Browser übertragen. Dies wurde nicht verfolgt, um den Fokus auf die E-Mail und die Formalisierung zu behalten. Die Bedeutung der E-Mail-Adresse in Bezug auf die Zuordnung einer Herkunft kann nicht geleugnet werden und ist essentiell. Eine kryptographisch gesicherte E-Mail-Adresse im `FROM`-Feld ohne Merkwürdigkeiten und `SENDER`-Feld bietet eine sehr gute Möglichkeit der Herleitung der Herkunft. Andererseits gibt es noch weitere mögliche Informationen, um die Herkunft einer E-Mail abzuleiten. Dazu erforderlich ist eine tiefergehende Beschäftigung mit der Herkunft einer E-Mail, welche für bereits vertraute Kontakte nicht sinnvoll ist.

Das grundsätzliche Problem an der Darstellung der E-Mail-Adresse ist, dass die Angreiferin die Darstellung frei wählen kann und damit menschliche Schwächen explizit ausnutzen kann. Die Herausforderung für die Darstellung ist die Unterscheidung zwischen Angriffen und legitimen Nachrichten. Der Übergang

[7] Ein weiterer Einstieg bietet folgender Blog-Eintrag: https://docs.microsoft.com/en-us/typography/develop/word-recognition

kann dabei fließend sein und führt zwangsläufig zu Fehlern bei der Klassifikation. Durch die Heuristiken ist eine Verallgemeinerung möglich und bietet eine Orientierung bei der Sicherheitsanalyse abseits von E-Mail-Adressen. Bei Verfahren, welche ohne Berücksichtigung der menschlichen Kognition entwickelt wurden, ist die direkte Anwendung der Formalisierung nicht ohne eigene Experimente möglich, aber sie bietet eine Hilfestellung. Statt die Unsicherheit der E-Mail-Adresse in eigenen Experimenten zu untersuchen, wird eher ein konstruktiver Ansatz verfolgt.

Das sichere Darstellungsverfahren im nächsten Kapitel bietet hierzu eine sinnvolle Ergänzung, um die bereits bekannte Personen oder Organisationen wiederzuerkennen.

Kapitel 6

Sicherheit durch Wiedererkennung

Dieses Kapitel umfasst die Konstruktion und Analyse eines Verfahrens basierend auf der Wiedererkennung der Herkunft einer Nachricht. Das Verfahren ist eine Demonstration einer konstruktiven Nutzung der Formalisierung und zeigt gleichzeitig, dass ein Fokus auf bisher vernachlässigte aber klassische UI-Elemente einer Anwendung gelegt wird. Bevor das Verfahren vorgestellt wird, werden zunächst der Hintergrund und verwandte Arbeiten zur Wiedererkennungsfähigkeit von Menschen dargestellt. Nach der Vorstellung des Verfahrens wird die Sicherheit analysiert und die praktische Umsetzung diskutiert.

6.1 Hintergrund und verwandte Arbeiten

In dem vorherigen Kapitel wurde bereits hervorgehoben, dass eine Darstellung auf Basis von Zeichenketten und der Interpretation dieser sicherheitskritische Fehler fördert. Studien zu Authentifizierungsverfahren zwischen Mensch und Computer deuten darauf hin, dass die Wiedererkennung von zufälligen Zeichenketten oder Zahlenfolgen für Menschen eine Herausforderung ist und somit eine Darstellung von zufälligen Zeichenketten ebenso wenig eine geeignete Lösung ist. Biddle et al. fassen bekannte Untersuchungen zu graphischen Mensch-Computer-Authentifizierungsverfahren zusammen [16]. Sie verweisen auf Studien aus den 1970er Jahren und schlussfolgern, dass Menschen Bilder besonders gut wiedererkennen können [16, 126, 95]. Ein bekanntes und häufig untersuchtes Verfahren ist PassFaces [36, 23]. Das Verfahren hat die gleiche Oberfläche wie ein typisches PIN-Verfahren, aber im Unterschied hierzu werden Gesichter statt Ziffern angezeigt. Eine Person muss sich somit statt Ziffern Gesichter merken, wiedererkennen und auswählen. In einer Studie von Brostoff und Sasse wurden die Unterschiede zwischen einem normalen Passwort und PassFaces untersucht. Bei den Passwörtern gab es eine Fehlerrate bei der Eingabe des Geheimnisses von $15,1\%$ und bei PassFaces von $4,9\%$ [23]. Dies zeigt, dass die Teilnehmenden deutlich besser die Gesichter wiedererkennen konnten. Davis et al. variierten PassFace und ersetzten die Gesichter mit allgemeinen Bildern. Die konzeptuelle Idee bei den allgemeinen Bildern war,

dass Teilnehmende sich aus verschiedenen Bildern eine Geschichte konstruieren [36]. In der Studie von Davis et al. war die Authentifizierung mittels Gesichtern deutlich häufiger erfolgreicher und insbesondere nach längerer Zeit ohne Nutzung vom Verfahren [36].

Mensch-Computer-Authentifizierungsverfahren

Im Kontext der Mensch-Computer-Authentifizierungsverfahren abseits von klassischen Passwörtern wurden viele unterschiedliche Varianten und Konzepte untersucht. Grafische Darstellungen eignen sich besser als ziffernbasierten Darstellungen und insbesondere Gesichter sind geeignet.

Die Herausforderung in Authentifizierungsverfahren, wie zum Beispiel der PIN oder PassFaces, ist, dass in mehreren Runden ein oder mehrere zuvor gelernte Elemente aus einer Menge von ähnlichen Elementen ausgewählt werden. Dieses Prinzip wird bei dem hier vorgestellten Verfahren gegen Phishing nicht angewendet und aus diesem Grund werden allgemeinere Erkenntnisse zur Wiedererkennung verwendet. Dies ermöglicht die Untersuchung des Ursprungs der Sicherheit und entspricht der Empfehlung von Wiese und Roth in Bezug auf die Untersuchung der Sicherheit von Authentifizierungsverfahren [148].

Zur Untersuchung der Wiedererkennungs- und Gedächtnisleistungen gibt es verschiedene Experimenttypen. Einerseits gibt es standardisierte Tests zur Wiedererkennung, die Wiedererkennung in bestimmten Szenarien, wie zum Beispiel die Beobachtung einer Straftat als Augenzeugin, und die Erkennung einer Wiederholung in einer großen Anzahl an zufälligen Bildern. Das jeweilige Experiment wird jeweils kurz vorgestellt und diskutiert.

Wechsler Gesichtstest

Stimulus

Standardisierte Farbfotos von menschlichen Gesichtern. Die Teilnehmenden kennen die Personen zu den Gesichtern nicht und haben damit keine persönliche Bezüge zu den Gesichtern oder den dargestellten Personen.

Ablauf

1. **Lernphase:** Teilnehmende sehen nacheinander die Gesichter (*Zielgesichter*) und haben die Aufgabe, diese sich zu merken.
2. **Sofortige Wiedererkennung:** Teilnehmende sehen nacheinander 48 Fotos von Gesichtern, wobei 24 davon die *Zielgesichter* sind und 24 neue unbekannte Gesichter sind. Die Teilnehmenden geben an, ob das dargestellte Gesicht ein *Zielgesicht* ist oder nicht.
3. **Wiedererkennung:** 30 Minuten später sehen die Teilnehmenden wieder 48 Fotos von Gesichtern, wobei 24 davon die *Zielgesichter* sind und 24 neue unbekannte Gesichter sind. Die unbekannten Gesichter haben die Teilnehmenden nicht in der vorherigen Phase gesehen. Die Teilnehmenden geben an, ob das dargestellte Gesicht ein *Zielgesicht* ist oder nicht.

Messung

Das Experiment ist ein Signal-Detection-Experiment und die 4 möglichen Ereignisse (*TP, FP, TN, FN*) werden gezählt.

Erkenntnisgewinn

Durch die standardisierte Auswahl an Gesichtern wird gemessen, wie gut eine Person sich Gesichter merken kann und die Leistung von verschiedenen Personen kann verglichen werden.

Der Wechsler Gesichtstest wird zum Beispiel zur Diagnose von bestimmten Krankheiten verwendet oder um die Auswirkungen von bestimmten Krankheiten zu messen. In anderen Studien wird der Test zum Beispiel genutzt, um Menschen nach ihrer Gesichtserkennungsleistung zu gruppieren und Korrelationen aufzuzeigen [90]. Durch die standardisierten Darstellungen mit den gleichen Gesichtern ist dieser Aufbau nicht geeignet, um die Sicherheit eines Verfahrens zu untersuchen. Der Untersuchungsgegenstand ist eher der Mensch als die Darstellung.

Eine praktische Anwendung für die Wiedererkennung ist die Glaubwürdigkeit von Zeugenaussagen, zum Beispiel im Gerichtsverfahren. Bei der Wiedererkennung von verdächtigen Personen gibt es grundsätzlich zwei Möglichkeiten:

1. n Personen werden gleichzeitig gezeigt und die verdächtige Person soll benannt werden.
2. n Personen werden nacheinander gezeigt und die verdächtige Person soll benannt werden.

Diese und weitere Fragestellungen wurden in Experimenten untersucht und der Aufbau der Experimente war ähnlich.

Augenzeugentest

Stimulus

1. Video von einer verdächtigen Aktion, zum Beispiel einem simulierten Diebstahl oder ein Foto einer Person
2. Fotos von Personen oder Gesichtern

Ablauf

1. **Beobachtungsphase:** Die Teilnehmenden sehen ein Video einer verdächtigen Person oder das Foto einer Person bzw. von einem Gesicht.
2. **Ablenkungsaufgabe:** Beispielsweise zehn Anagramme von Bundesstaaten der USA zu erkennen.
3. **Wiedererkennung:** Die Teilnehmenden sollen aus einer Menge von Darstellungen die verdächtige Person wiedererkennen.

Messung

Das Experiment ist ein Signal-Detection-Experiment und die 4 möglichen Ereignisse (*TP, FP, TN, FN*) werden gezählt [88].

Erkenntnisgewinn

Ein häufiger Untersuchungsgegenstand ist der genaue Ablauf vom Wiedererkennungsverfahren als A/B-Test. Beispielsweise die Instruktionen zur Wiedererkennung oder der Ablauf bei der Auswahl der verdächtigen Person. Dies betrifft vor allem die Frage, ob gleichzeitig die Auswahl dargestellt werden soll oder jede Person einzeln dargestellt wird.

Im Gegensatz zu dem Wechsler-Test wird hier das Verfahren untersucht. Damit ist das Ziel vom Experiment relativ ähnlich zu dem Ziel bei dieser Sicherheitsanalyse.

Die Studien für den Augenzeugentest finden unter Laborbedingungen statt. Im Gegensatz zu tatsächlichen Zeugen einer womöglich schweren Straftat sind die Teilnehmenden der Studien keiner solchen Stresssituation ausgesetzt. Dies ist eine natürliche Einschränkung der Studien, welche akzeptiert wird.

Der Augenzeugentest basiert auf einem Szenario. Ein Unterschied zu den Sicherheitsspielen und dem Anwendungsfall Nachrichtenempfang ist, dass der Augenzeugentest auf einem singulären Ereignis beruht. Im Gegensatz dazu versteckt sich ein Angriff eher im alltäglichen Kommunikationsstrom und bleibt unbemerkt. Die Darstellungen in einer Kommunikationsanwendung müssen über eine längere Zeit wiederholend gelernt werden.

Ein allgemeineres Forschungsgebiet ist das Wiedererkennen von Fotos und die Unterschiede zwischen

diesen.

Wiederholungserkennungssexperiment

Stimulus

Eine große Menge zufällig gewählter Bilder wird in Ziele, Aufmerksamkeitstestbilder und Lückenfüllerbilder unterteilt.

Ablauf

Die Studienteilnehmenden sehen eine Abfolge von einzelnen Fotos. Jedes Foto wird für ein bis zwei Sekunden dargestellt und zwischen jedem Foto ist eine kurze Pause (bis zu zwei Sekunden). Die Aufgabe der Teilnehmenden ist, eine Taste zu drücken, wenn sie der Meinung sind, ein Foto bereits gesehen haben. Ein Durchlauf einer Fotoreihe dauert etwa fünf Minuten. In manchen Experimenten konnten die Teilnehmenden mehrere Durchlaufe über unterschiedliche Fotos absolvieren.
Die Besonderheit ist, dass Zielfotos erst nach längerer Zeit (zum Beispiel nach 100 anderen Fotos) einmal erneut angezeigt werden. Fotos als Aufmerksamkeitstests werden nach maximal 7 anderen Fotos wieder dargestellt. Diese Fotos prüfen, ob die Person noch aufmerksam ist und die Aufgabe erledigen will. Die Füllerdarstellungen werden nur einmal dargestellt und wiederholen sich nicht.

Messung

Das Experiment ist ein Signal-Detection-Experiment und die 4 möglichen Ereignisse (*TP, FP, TN, FN*) werden gezählt.

Erkenntnisgewinn

Bei der Wiedererkennung können Unterschiede zwischen den Fotos erkannt werden und so kann untersucht werden, welche Fotos besser wiedererkannt und gemerkt werden.

Diese Wiederholungserkennungssexperimente eignen sich, um Darstellungen auszuwählen, welche besser wiedererkannt werden als andere. Ein Ziel hiervon ist es, Eigenschaften von Fotos für eine besondere gute Wiedererkennung zu finden und dies nach Möglichkeit mit Hilfe von Maschinellem Lernen oder Bildverarbeitung abzuschätzen. Damit können Fotos für eine bessere Wiedererkennung optimiert werden. Die Anwendungsfälle hierfür sind vielfältig und reichen von Material für Bildung hin zu Werbung, Spielen oder sozialen Medien [68].

In dem beschriebenen Experiment erfolgt aber keine Messung der Assoziation der Teilnehmenden mit dem Foto. Wenn eine fehlerhafte Erkennung einer Wiederholung erfolgte, dann wird aus der Messung nicht eindeutig, mit welchem Foto die Verwechselung erfolgte. Die implizite Annahme ist, dass dies mit dem fixierten Zielfoto verwechselt wurde.

Isola et al. haben das Wiederholungserkennungssexperiment mit unterschiedlichen Fotos durchgeführt und hatten 2.222 verschiedene Zielfotos und mehr als 8.000 andere Fotos [64]. Pro Bild lag die durch-

schnittliche korrekte Wiedererkennug bei $67,5\%$ (std: $13,6\%$) und die durchschnittliche falsche Wiedererkennung bei $10,7\%$ (std: $7,6\%$). Durchschnittlich haben 78 Teilnehmende ein Bild gesehen und Isola et al. schlussfolgern daraus, dass die Studienteilnehmenden ($n = 665$) nicht geraten haben [64]. Sie konnten beobachten, dass es aber starke Schwankungen zwischen den Bildern gab. Die Teilnehmenden wurden in zwei Gruppen geteilt und die besten 100 Bilder der ersten Gruppe hatten im Durchschnitt eine korrekte Wiederholungserkennungsrate von 93% und die zweite Gruppe hatte eine durchschnittliche korrekte Wiederholungserkennungsrate von 85% bei diesen Bildern. Korrelationstests und die Unterteilung in zufällige Gruppen der Teilnehmenden verstärkt dies [64].

Konsistent, aber schwankende Wiedererkennung

Die Wiedererkennung von Bildern ist konsistent zwischen verschiedenen Menschen, aber schwankt zwischen den Bildern.

Eine konsistente Wiedererkennung zwischen unterschiedlichen Personen ermöglicht es, einen nicht personalisierten Darstellungsraum auszuwählen. Die schwankende Wiedererkennung zwischen unterschiedlichen Bildern zeigt, dass der Darstellungsraum nicht einfach zufällig ausgewählt werden sollte, sondern eine geschickte Auswahl wichtig ist.

Isola et al. untersuchten Merkmale und Eigenschaften eines Fotos in Bezug auf eine Veränderung der Wiedererkennung. Einfache Merkmale, wie zum Beispiel der Farbraum, korrelierten nicht mit der Wiedererkennung. Die stärkste Korrelation wurde bei Fotos mit Personen gefunden [64]. Das bestätigt etablierte Theorien in der Kognitionswissenschaft und die Erkenntnisse aus den Forschungen zur Mensch-Computer-Authentifikation. Im Allgemeinen können Menschen Gesichter besser wiedererkennen als andere Objekte. Allerdings wurde in anderen Studien gezeigt, dass es bei anderen Objekten individuell ähnliche Wiedererkennungsleistungen gibt. Beispielsweise können Autoexpertinnen Autos besonders gut wiedererkennen [49].

Personen werden gut wiedererkannt

Im Allgemeinen können Menschen Bilder mit Personen oder Gesichtern besonders gut wieder erkennen.

Die Darstellung von Gesichtern ist im Kontext zur Darstellung der Herkunft einer Nachricht sinnvoll, weil häufig eine Person assoziiert wird. In diesem Abschnitt wird der Ansatz zur Darstellung der Herkunft durch Personen bzw. Gesichtern weiterverfolgt.

In dem Wiederholungserkennungssexperiment von Bainbridge et al. wurden als Stimulus 10.000 Gesichter verwendet und es wurde untersucht, ob es Unterschiede bei der Wiedererkennung zwischen den Gesichtern gibt [10]. Die Ergebnisse sind in Bezug auf die durchschnittliche korrekte Wiedererkennung schlechter als im Experiment von Isola et al. [64], aber in der Studie von Isola et al. [64] gab es ganz verschiedene Darstellungen aus unterschiedlichen Kontexten (Landschaften, Sportler, Büros, Supermärkte, Städte usw.) und dies erklärt eine bessere Erkennung einer Wiederholung. Bainbridge et al. konnten durch

eine Korrelationsanalyse zwischen zufällig gewählten Gruppen von Teilnehmenden eine Korrelation in Bezug auf die Wiedererkennung von Gesichtern zeigen. In weiteren Wiederholungserkennungssexperimenten verglichen Chapman et al. die Wiedererkennung bei bekannten und unbekannten Gesichtern [29]. Als bekannte Gesichter wurden Fotos von prominenten Personen gewählt. In einem weiteren Stimulus war die zweite Darstellung derselben Person ein anderes Foto. Die exakte Wiederholung des Fotos wurde in den Experimenten immer besser wiedererkannt. Die durchschnittliche korrekte Wiederholungserkennungsrate bei gleichen, aber bekannten Gesichtern betrug $73,8\%$ (std $= 20,6\%$) und bei gleichen, aber unbekannten Gesichtern betrug diese 35%(std $= 17,6\%$). Gleichzeitig lag die falsche Wiederholungserkennungsrate bei bekannten aber gleichen Gesichtern etwas höher als bei unbekannten aber gleichen Gesichtern ($11,2\%$ mit std $= 10,1\%$ gegenüber $9,9\%$ mit std $= 8,6\%$). In einer statistischen Analyse wurde die Hypothese, dass bekannte Gesichter besser korrekt wiedererkannt werden, bestätigt.

Bekannte Gesichter

Im Allgemeinen können Menschen Fotos von Gesichtern bekannter oder vertrauter Personen besser wiedererkennen.

Im Hinblick auf das zufällige Darstellungsverfahren kann dies bedeuten, dass in der Praxis die Gesichter bekannt und vertrauter werden. Ein Gesicht kann mit einer Identität verknüpft werden und dann besser wiedererkannt werden. Gleichzeitig ist das Gesicht einer neuen Herkunft unbekannt und damit erfolgt hier eher weniger eine Verwechselung.

In einer Studie von Goetschalckx et al. wurde das Wiederholungsentdeckungsexpierment in drei Phasen unterteilt und diese zeitlich auf einen Tag später und eine Woche später gestreckt [50]. Der Stimulus hat sich somit erst nach einer Woche erstmalig wiederholt. Dabei verringerte sich die durchschnittliche korrekte Wiedererkennung von 60% auf 47% nach einer Woche und die falsche Wiederkennungen erhöhten sich im Durchschnitt von 12% auf 27%. Dennoch war die Wiedererkennung bei den Teilnehmenden über die unterschiedlichen Zeitpunkte konsistent und korrelierten mit vorherigen Ergebnissen von Isola et al.[64]. In einer Zusammenfassung schlussfolgert Bainbridge, dass Erinnerungseffekte über längere Zeit anhalten [9].

Wiedererkennung nach längerer Zeit

Im Experiment von Goetschalckx et al. wurde deutlich, dass zwar die Wiederholungserkennungsrate nach längerer Zeit in absoluter Zahl sinkt, aber dennoch erstaunliche Leistungen erzielt. Die Effekte, welche in Experimenten mit kurzer Zeitspanne gemessen wurde, bleiben über eine längere Zeitspanne erhalten. Für einen Vergleich unterschiedlicher Stimuli eignen sich somit Experimente über eine kürzere Zeitspanne.

Die kognitiven Fähigkeiten wurden in Experimenten aus der Psychologie und Kognitionswissenschaft untersucht. Die bekannten Experimente reichen von Augenzeugen-Tests bis zu Wiederholungsexperimenten mit sehr vielen Gesichtern. In dem betrachteten Kontext erhalten Menschen Nachrichten von

unterschiedlichen Personen und ein Angriff soll in diesem Strom von Nachrichten untergehen. Ein erfolgreicher Angriff ist kein auffälliges singuläres Ereignis, wie zum Beispiel die Beobachtung von einem Raub, sondern einfach eine von vielen Nachrichten. Im Gegensatz zu den Wiederholungsexperimenten sehen Nutzerinnen die Darstellungen öfter über einen längeren Zeitraum und haben eine soziale Bindung zu den Personen, aber die Forschungsergebnisse in diesem Bereich deuten darauf hin, dass diese Faktoren zur Verbesserung beitragen.

Durch die vorherigen kognitiven Experimente und Erkenntnisse wird deutlich, dass Menschen die Wiederholung von Gesichtern erkennen und somit Gesichter ein geeigneter Darstellungsraum sind. Im Hinblick auf die Herkunft einer Nachricht ist ein Gesicht als Darstellung intuitiv und naheliegend.

6.2 Verfahren und Darstellungsraum

Das in diesem Kapitel betrachtete Verfahren basiert auf der Wiedererkennung von der Herkunft einer Nachricht. Jede neue Herkunft wird einem Symbol aus dem Darstellungsraum zugeordnet und dieses Symbol wird immer angezeigt. Die Herkunft einer Nachricht wird mit der (kryptographischen) Funktion Vrf eindeutig und korrekt bestimmt. Im Kontext der E-Mail kann dies beispielsweise eine digitale Signatur mittels DKIM, S/MIME oder PGP sein. Die Ausgabe dieser Funktion ist eine Zeichenkette. Statt diese Zeichenkette dem Menschen direkt zu zeigen, wird diese einem zufälligen Symbol zugeordnet. Sei $\mathcal{L}$ Darstellungsmenge mit n Elementen, $vrfy$ eine kryptographische Funktion zur Prüfung der Herkunft mittels digitaler Signaturen oder *Message Authentication Codes*. Das Verfahren 1 ist die abstrakte Formalisierung für n unterschiedliche Herkunftsbezeichnungen.

Gen	$\mathrm{R}_S(m)$
/ Create empty dict	/ Verify sender
$T \leftarrow \{\}$	$o \leftarrow vrf(m)$
$S \leftarrow \pi(\mathcal{L})$	**if** $o \in T$:
$i \leftarrow 0$	**return** $T[o]$
$n \leftarrow \lvert S \rvert$	**if** $o = \perp \vee i = n$:
return T, S, i, n	**return** $ERROR$
	$s \leftarrow S_i$
	$i \leftarrow i + 1$
	$T[o] = s$
	return s

Darstellungsverfahren 1: In dieser Darstellung wird jeder Herkunft eine zufällige Darstellung zugeordnet, wobei der Zustand Σ sich in diesem Verfahren aus T, S, i, n zusammensetzt und intern gespeichert wird. Auf eine Übergabe von diesem Zustand wird verzichtet.

Im vorherigen Abschnitt wurde gezeigt, dass Menschen besonders gut Gesichter wiedererkennen können. Die Gesichter aus dem *10kFaces*-Datensatz [8] bieten eine Grundlage für einen Darstellungsraum. Die-

ser kann weiter eingeschränkt bzw. besonders sortiert werden, so dass zunächst besonders vorteilhafte Darstellungen ausgewählt werden.

In diesem Verfahren kann die Angreiferin die Darstellung nicht stark beeinflussen. Die Darstellung ist aus der Angriffsperspektive quasi-zufällig, weil die Angreiferin nicht weiß, welches Gesicht ihr zugeordnet wird. Die Herausforderung für Nutzerinnen ist in diesem Verfahren nicht mehr die Interpretation der Darstellung, sondern in der Regel soll die Darstellung wiedererkannt werden und so eine Herkunft zugeordnet werden. Die Interpretation der Herkunft erfolgt erst im zweiten Schritt, wenn die Darstellung unbekannt ist und nicht wiedererkannt wird. Dies kann in einer dedizierten zusätzlichen Ansicht mit weiteren Informationen erfolgen und wird für dieses Verfahren nicht betrachtet. Für die Untersuchung der Sicherheit muss zunächst die Fehlerrate bei der Wiedererkennung einer Darstellung betrachtet werden.

6.3 Einbettung im Sicherheitsspiel

Zur Analyse der Sicherheit wird das Verfahren im formalen Sicherheitsspiel 4 (Origin) betrachtet. Das Spiel wurde in Kapitel 4 genauer erläutert und wird neben dem neuen Kognitionsspiel 2 wiederholt.

$\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}$

1 : $\Sigma^0_{\text{R}_S} \leftarrow \text{Gen}()$
2 : $\Sigma^\star_{\text{R}_S}, m_1 \leftarrow \text{Setup}^{\mathcal{H}^{\text{recognize}}}(\Sigma^0_{\text{R}_S})$
3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{r} \leftarrow \mathcal{H}^{\text{recognize}}(\text{R}_S(m_b, \Sigma^\star_{\text{R}_S}))$
6 : **return** $\hat{r} = 1, b$

$\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}_S}$

1 : $\Sigma^0_{\text{R}_S} \leftarrow \text{Gen}()$
2 : $\Sigma^\star_{\text{R}_S}, m_1, \varphi \leftarrow \text{Setup}^{\mathcal{H}^{\text{origin}}}(\Sigma^0_{\text{R}_S})$
3 : $m_0 \leftarrow\!\!\$\ \mathcal{A}(\text{Leak}(m_1))$
4 : $b \leftarrow\!\!\$\ \{0,1\}$
5 : $\hat{\varphi} \leftarrow \mathcal{H}^{\text{origin}}(\text{R}_S(m_b, \Sigma^\star_{\text{R}_S}))$
6 : **return** $\varphi = \hat{\varphi}, b$

Kognitionsspiel 2: Auf der linken Seite ist das neue Wiederholungserkennungsspiel darstellt. Es orientiert sich sehr an dem Sicherheitsspiel Origin, welches auf der rechten Seite wiederholt dargestellt ist.

Unter der Annahme, dass keine Interpretation der Darstellung erfolgt, ist der Unterschied zwischen beiden Spielen die Ergänzung der Herkunft aus der Darstellung.

Ergänzend zum Sicherheitsspiel wird das Kognitionsspiel 2 formuliert. In diesem Spiel geben die Person an, ob eine Herkunft wiederholt vorkommt oder nicht. Diese Anfrage wird mit recognize an $\mathcal{H}$ gestellt. In diesem Spiel bedeutet die Ausgabe $1, b$, dass die Herkunft von m_b wiederholt vorgekommen ist. Im Gegensatz zum ersten Sicherheitsspiel wird die Herkunft nicht verglichen.

Durch die zufällig gewählte Darstellung in dem Verfahren kann die Darstellung nicht interpretiert werden, sondern der Unterschied zwischen beiden Spielen ist die Zuordnung zu einer bekannten Herkunft. Diese Zuordnung einer Herkunft kann nur erfolgen, wenn die Darstellung als bekannt wahrgenommen wird. Die Zuordnung als Anfrage an $\mathcal{H}$ wird mit map modelliert und simuliert eine Gedächtnisabfrage. Die Rückgabe ist eine Herkunft oder keine Herkunft (also unbekannt). Unter der Annahme, dass die

Darstellungen nicht interpretiert werden, ist die Lücke zwischen beiden Spielen die Zuordnung einer Darstellung zu einer Herkunft und die Beziehung kann für $b \in \{0,1\}$ wie folgt dargestellt werden. Das Ereignis $\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}_S}(1,b)$ entspricht dem Ereignis $\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}(1,b)$ und dem Ereignis $\mathcal{H}^{\text{map}}(\text{R}_S(m_b, \Sigma^{\star}_{\text{R}_S})) = \varphi$.

Das Ereignis $\text{Origin}^{\mathcal{A},\mathcal{H}^{\text{origin}}}_{\text{Setup,Leak,Gen,R}_S}(0,b)$ entspricht dem Ereignis $\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}(0,b)$ oder dem Ereignis $\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}(1,b)$ und dem Ereignis $\mathcal{H}^{\text{map}}(\text{R}_S(m_b, \Sigma^{\star}_{\text{R}_S})) \neq \varphi$.

Dabei ist zu berücksichtigen, dass bei einer Zurückweisung ($\hat{r} = 0$) im $\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}$ Spiel die Darstellung unbekannt ist und somit keiner bekannten Identität zugeordnet wird. Bei einer Zurückweisung kann somit kein Zuordnungsfehler in Bezug auf die Identität im Origin Spiel auftreten und die Identitäten sind verschieden. Erst bei einer späteren Darstellung oder unter Bezug von anderen UI-Elementen kann diese erfolgen. Wenn eine Herkunft als Wiederholung erkannt wird, kann diese korrekterweise oder fälschlicherweise der Herkunft φ zugeordnet werden.

Das Ereignis $\varphi = \hat{\varphi} \wedge b = 0$ aus Origin entspricht dem Ereignis die Herkunft vom Angriff ist unbekannt, $\hat{r} = 1 \wedge b = 0$ aus $\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}$, die Herkunft φ zugeordnet wird oder dem Ereignis $\hat{r} = 1 \wedge b = 1$, die Angriffsherkunft bekannt ist und die Herkunft φ zugeordnet wird. Diese Ereignisse werden später genauer betrachtet, aber zunächst wird der Bezug zwischen dem Wiederholungsexperiment und dem $\text{Recog}^{\mathcal{A},\mathcal{H}^{\text{recognize}}}_{\text{Setup,Leak,Gen,R}_S}$-Spiel dargestellt. Dieses unterscheidet sich vom vorgestellten Wiederholungsexperiment von Isola et al. [64] Die offensichtlichen Unterschiede sind die Abwesenheit von einer Angreiferin im Experiment und die experimentelle Bestimmung der Erkennungsrate einer Wiederholung im Wiederholungsexperiment gegenüber der abstrakten Betrachtung der Erfolgswahrscheinlichkeit eines Angriffs. Im ersten Schritt wird zunächst das Wiederholungsexperiment formalisiert und schrittweise durch die Gestaltung neuer abstrakter Experimente umgeformt.

Die Beziehung zwischen dem Experiment und dem Spiel wird verdeutlicht. Das ist kein klassischer mathematischer Beweis, sondern eine Argumentationskette zur Abschätzung und Plausibilität der Sicherheit. Für die Formalisierung vom Wiederholungsexperiment wird auf die Einbeziehung des Aufmerksamkeitstests, also die Wiederholung einer Grafik nach kurzer Zeit (bis zu sieben Fotos später), verzichtet. Dieser Teil des Experiments dient nur zur Kontrolle, ob die Personen noch aktiv teilnehmen, zur Erinnerung an die Aufgabe im Experiment (Drücken einer Taste bei einer Wiederholung), als eine Ablenkungsaufgabe für die relevanten Wiederholungen der Grafiken und als positives Feedback für die teilnehmenden Personen. Für die Formalisierung wird das Experiment auf eine einfache Form reduziert. Dabei ist $\mathcal{L}$ der Darstellungsraum, l die untere Schranke an Bildern nachdem eine Wiederholung erfolgt und u die obere Grenze, nach der eine Wiederholung erfolgen soll.

Eine Vereinfachung vom Wiederholungsexperiment von Isola et al. [64] wird in Kognitionsspiel 3 dargestellt. In diesem Kognitionsspiel wird eine zufällige Permutation über den Darstellungsraum gewählt, der Index (i) vom Wiederholungselement (S_i) zufällig gewählt (in dem tatsächlichen Experiment wurde dies vorab für alle Personen zufällig gewählt). Es wird so die Anzahl der Bilder bis zur Wiederholung gewählt. In den tatsächlichen Experimenten waren dies 91 bis 109 Bilder. Im Anschluss werden der Per-

RepeatedRecog$_{\mathcal{L},l,u}$

1 : $S \leftarrow \pi(\mathcal{L})$
2 : $n \leftarrow |S|$
3 : $i \leftarrow\!\!\$\ \{l, \ldots, u\}$
4 : $r_0 \leftarrow \mathcal{H}^{\text{recognize}}(S_i)$
5 : **for** $j \in \{1 \ldots n\}$:
6 : $\quad r_j \leftarrow \mathcal{H}^{\text{recognize}}(S_j)$
7 : **return** $r_0 \ldots r_n, i, s_i$

Kognitionsspiel 3: Dieses Kognitionsspiel ist eine Vereinfachung vom Wiederholungsexperiment von Isola et al. [64].

	$j = i$	$j \neq i$
$r = 1$	richtiger Treffer	falscher Treffer
$r = 0$	falsche Zurückweisung	korrekte Zurückweisung

Tabelle 6.1: Mögliche Ereignisse aus dem Kognitionsspiel 3. $r = 1$ bedeutet, dass die Person eine Wiederholung meldet. $r = 0$ bedeutet, dass die Person keine Wiederholung meldet. Die tatsächliche Wiederholung erfolgt an Position i.

son die Bilder angezeigt und die Person gibt per Tastendruck an, ob eine Wiederholung erfolgte. Diese Rückmeldung wird mit r bezeichnet, wobei es folgende Bedeutungen hat:

1. $r = 1$: Die Taste wurde gedrückt und die Person meldet eine Wiederholung.
2. $r = 0$: Die Taste wurde nicht gedrückt und die Person meldet keine Wiederholung.

In der einfachen Form des Experiments wird nur eine Wiederholung betrachtet. Im tatsächlichen Experiment haben die Teilnehmenden 120 Bilder in einem Level (also einem Teilexperiment) gesehen. Es ist also plausibel, dass die Teilnehmenden in einem Level mehrere Bilder wiedererkennen mussten, aber die Beschreibung des Experimentes ist hierzu ungenau [10, 64]. Für die Überleitung zum Sicherheitsspiel ist die Annahme von nur einer Wiedererkennung nötig und die möglichen Ergebnisse in dem Kognitionsspiel werden im Folgenden betrachtet. Sei i der Index der Wiederholung und sei S die zufällige Permutation der Darstellungen aus $\mathcal{L}$, wobei $S_i \in S$ das Wiederholungselement ist und vereinfacht an erster Stelle steht. Tabelle 6.1 stellt die möglichen Ereignisse für Darstellung S_i dar, wobei $j, i \in \{1 \ldots n\}$ und j der Index der Wiederholung ist.

In diesem Kognitionsspiel wird die korrekte Wiederholungserkennungsrate bzw. die falsche Wiederholungserkennungsrate von dem Element S_i bestimmt. Dabei wird kein Angriff berücksichtigt und im Gegensatz zum Sicherheitsspiel wird nicht das Ergebnis einer bestimmten Darstellung berücksichtigt. Aus diesem Grund wird das Kognitionsspiel angepasst, indem zufällig die Rückgabe der korrekten Wiederholung oder der Darstellung vor der korrekten Wiederholung erfolgt.

In dem Kognitionsspiel 4 wird durch einen zufälligen Münzwurf entschieden, welche Rückgabe ausgegeben wird. Für $\mathcal{H}$ gibt es hierzu keinen Unterschied. Durch die zufällige Permutation ist die Wahl von

SimulatedAttackRecog$_{\mathcal{L},l,u}$

1 : $S \leftarrow \pi(\mathcal{L})$
2 : $n \leftarrow |S|$
3 : $i \leftarrow_{\$} \{l, \ldots, u\}$
4 : $r_0 \leftarrow \mathcal{H}^{\text{recognize}}(S_i)$
5 : **for** $j \in \{1 \ldots n\}$:
6 : $\quad r_j \leftarrow \mathcal{H}^{\text{recognize}}(S_j)$
7 : $b \leftarrow_{\$} \{0, 1\}$
8 : **if** $b = 0$:
9 : $\quad$ **return** $r_{i-1}, 0$
10 : **return** $r_i, 1$

Kognitionsspiel 4: In dem neuen Kognitionspiel wird das Sicherheitsspiel 2 mit einem Angriff simuliert. Dies ist nur möglich, weil $\mathcal{A}$ nicht die Darstellung kontrollieren kann, sondern eine zufällige Darstellung erhält.

der Darstellung an Position $i-1$ zufällig und die Fehlerwahrscheinlichkeit kann mit den entsprechenden Fehlerraten abgeschätzt werden.

Beim Übergang zwischen dem Kognitionsspiel 4 und dem Spiel 2 ist eine Betrachtung vom Verfahren nötig. In $\mathrm{Gen}()$ wird eine zufällige Permutation von $\mathcal{L}$ (also $\pi(\mathcal{L})$) erzeugt und die Darstellung an Position i aus dem Kognitionsspiel ist die Darstellung von m_1 aus dem Sicherheitsspiel. Die gezeigten Darstellungen von 1 bis $i-2$ aus dem Kognitionsspiel 4 und die Darstellung am Anfang von S_i bilden Setup im Spiel 2 und stellen den Kontext vom Sicherheitsspiel dar. Dieser kann auf das Sicherheitsspiel Origin unter Berücksichtigung der Abschätzung übertragen werden. $\mathrm{Leak}(m_1)$ kann m_1 umfassen unter der Annahme, dass $vrfy$ auf einem Signaturverfahren (z.B. DKIM, S/MIME, PGP) basiert. Damit kann der Fall $T[vrf(m_0)] = T[vrf(m_1)]$ nicht auftreten. Gleichzeitig ist die Wahl der Darstellung einer Herkunft zufällig und kann nicht von $\mathcal{A}$ beeinflusst werden. Die Simulation eines Angriffs erfolgt mittels der Darstellung an Position $i-1$ im Experiment. Die Ausgabe von $\mathcal{H}^{\text{recognize}}(S[i-1])$ im Experiment entspricht der Ausgabe von $\mathcal{H}^{\text{recognize}}(m_0, \Sigma^{\star}_{\mathrm{R}_S})$ und $\mathcal{H}^{\text{recognize}}(S[i])$ entspricht $\mathcal{H}^{\text{recognize}}(m_1, \Sigma^{\star}_{\mathrm{R}_S})$ Damit entspricht die Rückgabe aus dem Kognitionsspiel 4 der Rückgabe aus dem Spiel 2.

Ableitung der Sicherheit

Aus der Konstruktion des Verfahrens leitet sich ab, dass die korrekte Wiederholungserkennungsrate und die falsche Wiederholungserkennungsrate ein essenzieller Bestandteil der Sicherheitsanalyse für das Verfahren sind.

Eine Einschränkung bei der Sicherheitsanalyse ist, dass Setup sich aus dem Experimentaufbau ableitet. Dieses ist damit fixiert und nicht allgemein gültig. Aus diesem Grund ist diese Einbettung kein klassischen Beweis, sondern eine transparente und nachvollziehbare Herleitung der Sicherheit. Gleichzeitig erscheint das Setup sehr streng zu sein, denn eine wirkliche Lernphase der bekannten Darstellungen findet nicht statt. Ein Angriff auf eine Person/Organisation, nachdem diese nur einmal gesehen wurde, ist

ein ungewöhnlicher Phishing-Angriff. Aus dieser Perspektive ist die Sicherheitsanalyse eine konservative Betrachtung. Im nächsten Abschnitt werden die Möglichkeiten zur Abschätzung der korrekten Erkennungsrate und der falschen Erkennungsrate betrachtet. Es wird ein Modell aufbauend auf maschinellem Lernen aus den Experimentaldaten vorgestellt. Dies ist nötig, um die Sicherheitsanalyse fortzusetzen.

6.4 Abschätzung und Optimierung der Fehler

Die Sicherheitsabschätzung basiert auf den kognitiven Fähigkeiten der Nutzerinnen bei der Wiedererkennung von den dargestellten Gesichtern und gleichzeitig optimiert es darauf die Auswahl und die Reihenfolge der Gesichter.

Aufbauend auf dem Wiederholungsexperiment wurden in der Vergangenheit Modelle mittels Maschinellen Lernens entwickelt und veröffentlicht. Dieser Ansatz wird für den Zweck hier verwendet. Es wird zunächst ein Modell ausgewählt und dieses reproduziert sowie validiert. Im Gegensatz zu bisherigen Modellen muss ein Modell zur falschen Erkennung einer Wiederholung entwickelt werden. Mit diesem Modell werden alle Darstellungen in Bezug auf falsche und korrekte Erkennung einer Wiederholung abgeschätzt. Je nach Anzahl der möglichen Darstellungen kann die Sicherheit abgeschätzt werden. Im Folgenden wird eine falsche Erkennung einer Wiederholung als *false positive* (FP) und eine korrekte Erkennung einer Wiederholung als Treffer (H, *true positive, hit*) abgekürzt. Die jeweilige Erkennungsrate wird mit *false positive rate* oder abgekürzt mit FPR bzw. mit *true positive rate* oder abgekürzt mit TPR bezeichnet.

6.4.1 Beschreibung vom Datensatz

Der Datensatz *10kFaces* wurde von Bainbridge et al. erstellt. Zur Erstellung wurde nach Namen aus dem US-Census-Datensatz von 1990 mit einer Internetsuchmaschine gesucht und die Gesichter heruntergeladen. In einer manuellen Durchsicht wurden die Bilder gefiltert. [8, 10] Die Auswahl der Gesichter folgt den demographischen Faktoren der Verteilung vom US-Census-Datensatz von 1990 und umfasst 10.167 Gesichter [10]. Dieser Datensatz ist mit dem US-Urheberrecht vereinbar und wurde auf Anfrage[1] von Bainbridge et al. zur Verfügung gestellt.

Bainbridge et al. nutzen diesen Datensatz in einem Wiederholungsexperiment [10]. Dieses Experiment wurde mittels der M-Turk-Plattform von Amazon durchgeführt und 877 Personen nahmen teil. In diesem Experiment haben Teilnehmende innerhalb von 4,8 Minuten 120 Fotos nacheinander betrachtet und mussten angeben, ob ein Bild sich wiederholt hat. Eine Wiederholung erfolgte nach 91 bis 109 Bildern. Dies umfasst ein *Level* und eine Person kann an mehreren *Levels* teilnehmen, wobei es keine Schwierigkeitsunterschiede zwischen den *Levels* gibt. Durchschnittlich wurde jedes Gesicht von $81,7$ unterschiedlichen Personen betrachtet. Im Durchschnitt lag die HR bei $51,6\%$ bei einer Standardabweichung von $12,6\%$,

[1] https://www.wilmabainbridge.com/facememorability2.html

Modell	Wiederholungserkennungsrate	Korrigierte Wiederholungserkennungsrate
VGG16	0,445	0,579
ResNet50	0,433	0,607
SENet50	0,448	0,601
ResVGG	0,423	0,626
SENRes	0,452	0,631
SENVGG	0,468	0,605
SENResVGG	0,445	0,634
Mensch [10]	0,68	

Tabelle 6.2: Spearman Rankkorrelationkoeffizient bei den Modellen von Younesi und Mohsenzadeh [150] und im Vergleich dazu dem Mittelwert zwischen den menschlichen Gruppen aus dem ursprünglichen Experiment [10]. Für korrigierte Wiedererkennungsrate wurde von der Wiedererkennungsrate die falsche Wiedererkennungsrate abgezogen. SENRES, SENVGG, SENResVGG sind die Kombination aus den jeweiligen einzelnen Modellen.

wobei die Werte zwischen 15,5% und 89,9% schwankten. Die FAR betrug im Durchschnitt $14,4\%$ bei einer Standardabweichung von $8,7\%$, wobei die Werte zwischen 0% und $51,5\%$ schwankten. Bainbridge et al. untersuchten die Korrelation bei den Erkennungsraten zwischen den Teilnehmenden. Dazu wurden die Teilnehmenden wiederholt zufällig in zwei Gruppen unterteilt und die Korrelation der Erkennungsraten zwischen den beiden Gruppen betrachtet [10]. Die Korrelation wurde mittels Spearman Rankkorrelationkoeffizient bestimmt. Es wurden mehr als 25 dieser Zufallsgruppeneinteilungsversuche durchgeführt. Der durchschnittliche Korrelationswert bei der korrekten Wiederkennung betrug 0,68 (Minimum 0,66, Maximum 0,69) und bei der falschen Wiedererkennung betrug der Durchschnittswert 0,69 (Minimum 0,67, Maximum 0,71) [10]. Sie schlussfolgern, dass die korrekte beziehungsweise falsche Wiedererkennung zwischen Gesichtern schwankt, aber es in Bezug auf die Betrachtenden gleiche Tendenzen gibt. Die korrekte Wiederholungserkennungsrate und die falsche Erkennungsrate wurde nicht für alle Gesichter aus dem Datensatz bestimmt. Dies wird aber benötigt und deswegen abgeschätzt.

6.4.2 Modellauswahl

Aufbauend auf den Ergebnissen der Wiederholungsexperimente wurden Modelle mittels Maschinellem Lernen zum Vorhersagen der Wiedererkennung eines Bildes erstellt. Bylinskii et al. vergleichen unterschiedliche Modelle auf verschiedenen Datensätzen, aber nicht in Bezug auf den *10KFaces*-Datensatz [25]. Younesi und Mohsenzadeh haben Modelle auf Basis vom *10KFaces*-Datensatz vorgeschlagen und untersucht [150]. Ihre Modelle basieren auf den vortrainierten Modellen SeNet50, ResNet50 sowie VGG16 und somit auf einem Datensatz zur Gesichtserkennung [124]. Bei den beiden Modelle wurden nur die korrekte Wiederholungserkennungsrate und die korrigierte Wiederholungserkennungsrate (TPR - FPR) berücksichtigt. Tabelle 6.2 ermöglicht einen Vergleich der Modelle mittels des Spearman Rankkorrelationkoeffizienten zwischen der Vorsage und dem tatsächlichen Ergebnis.

Hierbei ist zu erkennen, dass die Ergebnisse bei der Wiederholungserkennungsrate relativ ähnlich sind,

wobei es bei der korrigierten Wiederholungserkennungsrate Unterschiede gibt. Im Vergleich zu der Korrelation zwischen menschlichen Gruppen wird deutlich, dass es noch Verbesserungspotential gibt. Das Ziel der nachfolgenden Analyse ist aber nicht die Verbesserung der Modelle, sondern die Nutzung dieser Modelle in einem Anwendungsfall und die Validierung. Dies bildet somit eine weitere Motivation zur Verbesserung der Modelle zur Vorhersage und damit zu einer genaueren Abschätzung der Sicherheit von dem Verfahren.

Younesi und Mohsenzadeh haben ihre Python-Programme öffentlich zur Verfügung gestellt [2]. Die Programme wurden angepasst und um die Vorhersage der falschen Wiedererkennnung ergänzt. Die Konfiguration der Modelle wurde dabei nicht verändert, sondern wie publiziert übernommen.[3] Für eine bessere Vergleichbarkeit der Ergebnisse wurde anfangs der Datensatz in einen Trainings- und Testdatensatz unterteilt. Der Testdatensatz wurde wiederum in einen Datensatz zur Validierung während des Trainings und einem abschließenden Testdatensatz unterteilt. Beim Trainieren der jeweiligen Modelle wurden somit immer die gleichen Datensätze verwendet. Der Aufbau der Modelle wurde von Younesi und Mohsenzadeh übernommen, eine Beschreibung und Erläuterung der Modelle ist dort zu finden [150].

Zur Bestimmung der korrekten und falschen Wiederholungserkennungsrate wird jeweils das beste Modell nach den Spearman Rankkorrelationkoeffizienten ausgewählt. Dieses Modell wird zur Bestimmung der korrekten bzw. falschen Wiederholungserkennungsrate für alle Darstellungen verwendet. Danach werden Teilmengen der Gesichter nach den Wiederholungserkennungsraten betrachtet.

6.4.3 Ergebnis

Modellwahl

Tabelle 6.3 zeigt die Ergebnisse der reproduzierten Modelle. Die Korrelation der Modelle ResNet50 und SeNet50 in Bezug auf die korrekte Wiedererkennung sind sehr ähnlich zu den Ergebnissen von Younesi und Mohsenzadeh. Einzig das Modell basierend auf VGG16 war deutlich schlechter als von Younesi und Mohsenzadeh publiziert.

Die gelernten Modelle zur falschen Wiedererkennung haben eine deutlich höhere Korrelation. Dies entspricht den Erwartungen, denn die Korrelationen in Bezug auf korrigierte Wiedererkennung waren bei Younesi und Mohsenzadeh deutlich besser. Dies wird beim Vergleich des mittleren quadratischen Fehlers (MSE) der Modelle ebenso deutlich. Der mittlere quadratische Fehler ist bei allen Modellen zur Vorhersage der korrekten Wiederholungserkennungsrate sehr ähnlich und der falschen Wiederholungserkennungsrate sogar gleich. Der mittlere quadratische Fehler bei den Modellen zur korrekten Wiedererkennung ist aber deutlich höher als bei der falschen Wiederholungserkennungsrate.

[2] https://github.com/mamyou96/FaceMemNet

[3] Der Programmcode zum Lernen und zur Auswertung ist unter http://dx.doi.org/10.17169/refubium-41698 erreichbar.

MSE	Spearman Koef.	Art	Model
0,014	**0,451**	**HR**	**senet50**
0,014	0,431	HR	resnet50
0,015	0,319	HR	vgg16
0,005	0,536	FAR	senet50
0,005	**0,571**	**FAR**	**resnet50**
0,005	0,532	FAR	vgg16

Tabelle 6.3: Es werden die unterschiedlichen gelernten Modelle verglichen. Bei der Vorhersage von einer falschen Wiederkennung (FAR) schneiden die Modelle deutlich besser ab als bei der Vorhersage einer korrekten Wiedererkennung (HR). Innerhalb einer Art der Vorhersage sind die Unterschiede deutlich geringer. Die hervorgehobenen Modelle wurden ausgewählt.

Min	Mittelwert	Std	Max	n
-0.019	0.013	0.011	0.025	100
-0.019	0.046	0.015	0.064	1000
-0.019	0.089	0.028	0.131	5000
-0.019	0.138	0.060	0.306	10000
-0.019	0.141	0.065	0.418	10167

Tabelle 6.4: Die Vorhersage der falschen Wiederkennungsrate (FAR) bei den besten n Gesichtern. Die negativen Werte sind Vorhersagefehler und verdeutlichen die Einschränkung der Modelle als ein Schätzer.

Für die nachfolgende Betrachtung wurde das SeNet50-Modell zur Vorhersage der korrekten Wiedererkennung und das ResNet50-Modell zur Vorhersage der falschen Wiedererkennung verwendet.

FPR Vorhersage

Die minimale Falsche-Wiederholungserkennungsrate (FPR) beträgt $-0,019$ und der maximale Wert beträgt $0,431$, wobei der Durchschnittswert bei 10.168 Gesichtern $0,141$ bei einer Standardabweichung von $0,065$ beträgt. Die negativen Werte sind Fehler bei der Schätzung, denn negative Werten sind außerhalb vom Definitionsbereich. Die Abbildung 6.1 illustriert die Verteilung der geschätzten falschen Wiederholungserkennungsrate und Tabelle 6.4 zeigt die deskriptiven Kennzahlen für eine Auswahl der ersten n aufsteigend sortierten Gesichter. Bei 5.000 Gesichtern beträgt die durchschnittliche falsche Wiederholungserkennungsrate unter 10% und die maximale falsche Wiederholungserkennungsrate bei $0,131$.

TPR Vorhersage

Die minimale korrekte Trefferrate beträgt $0,202$ und der maximale Wert beträgt $0,902$, wobei der Durchschnittswert bei 10.168 Gesichtern $0,513$ bei einer Standardabweichung von $0,09$ beträgt. Die Abbildung 6.1 illustrierte die Verteilung der geschätzten korrekten Wiederholungserkennungsrate und Tabelle 6.5 zeigt die deskriptiven Kennzahlen für die Auswahl der ersten absteigend sortierten n Gesichter. Bei den besten 5.000 Elementen liegt die durchschnittliche korrekte Trefferrate und die minimale korrekte

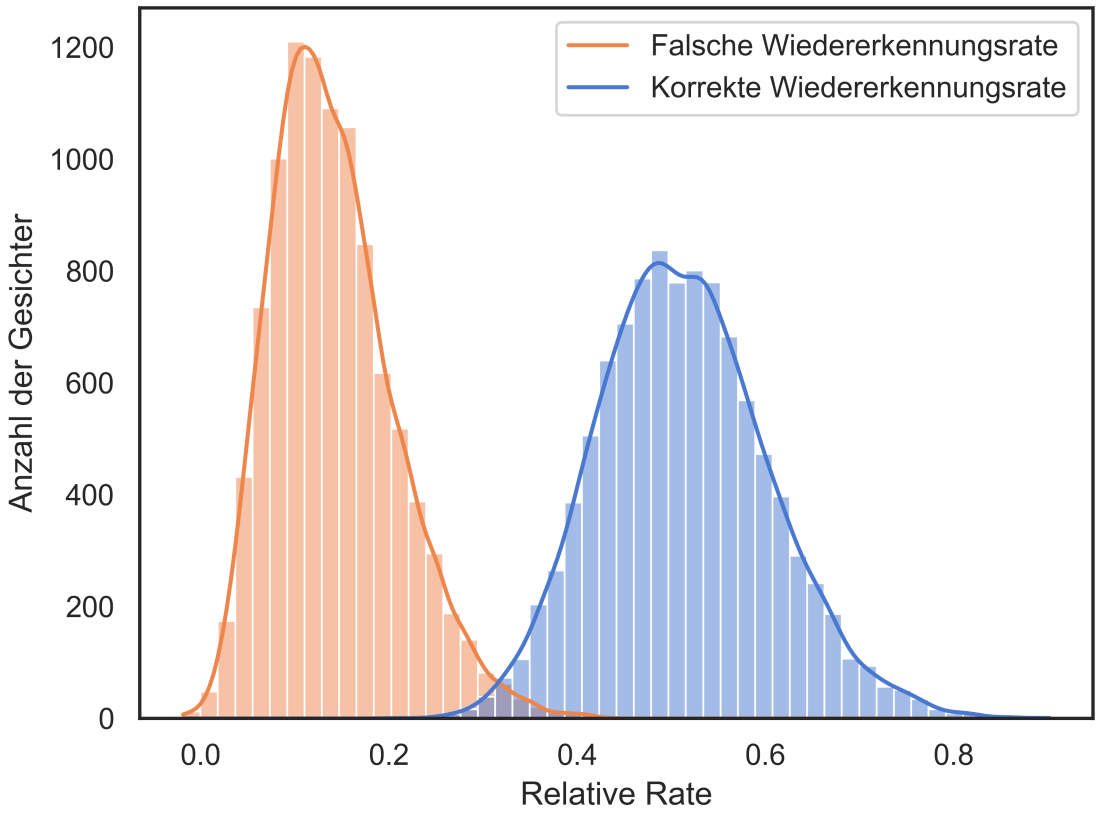

Abbildung 6.1: Verteilung in absoluten Werten der geschätzten korrekten Wiederholungserkennungsrate (HR) und der falschen Wiederholungserkennungsrate (FAR).

Min	Mittelwert	Std	Max	n
0.746	0.780	0.033	0.902	100
0.633	0.683	0.044	0.902	1000
0.510	0.586	0.060	0.902	5000
0.335	0.517	0.087	0.902	10000
0.231	0.513	0.090	0.902	10167

Tabelle 6.5: Die Vorhersage der korrekten Wiederkennungsrate (HR) bei den besten n Gesichtern.

Trefferrate bei über 50%.

6.4.4 Diskussion

Modelle

Es wurden mehrere verschiedene Modelle reproduziert. Die Modelle mit Ausnahme vom VGG16-Modell zur Vorhersage der korrekten Wiedererkennung sind vergleichbar mit den bekannten Modellen, aber es gibt im Vergleich zu der Korrelation zwischen Menschengruppen noch Verbesserungspotential. Das Ziel der Arbeit war aber nicht die Verbesserung der Modelle, sondern die Reproduktion und Nutzung der Modelle. Für die Nutzung des Modells wurde das beste ausgewählt, also SeNet50. Für die Modelle zur Vorhersage der falschen Wiederholungserkennungsrate fehlt der menschliche Referenzwert und dieser kann aus den akkumulierten zur Verfügung gestellten Daten nicht rekonstruiert werden. Absolut betrachtet sind diese Modelle aber deutlich besser als die Modelle zur Vorhersage der korrekten Wiedererkennung. Für die Vorhersage der falschen Wiedererkennung wurde ResNet50 ausgewählt. Die Hoffnung, dass die Modelle auch zum Lernen der falschen Wiedererkennung verwendet werden können, wurde bestätigt.

Vorhersage

Diese Kennzahlen zur falsche Wiedererkennung und zur korrekten Wiedererkennung sind vergleichbar mit den experimentell bestimmten Daten. Die Daten für das Experiment sind eine zufällige Teilmenge und damit ist die Verteilung ähnlich. Aus diesem Grund scheinen die bestimmten Werte über den gesamten Datensatz plausibel zu sein und die Fehler im Testset waren gering. Sowohl bei der korrekten Wiedererkennung als auch bei der falschen Wiedererkennung gibt es deutliche Unterschiede zwischen den Gesichtern aus dem Datensatz. Die Differenz zwischen dem maximalen und minimalen Werten ist bei der falschen Wiedererkennung bei über 40% und bei der korrekten Wiedererkennung sogar bei fast 70%. Die Wahl der Darstellungen aus dem Datensatz hat somit einen großen Einfluss auf die Sicherheit.

In Bezug auf die falsche Wiedererkennung gibt es eine Auswahl von 5.000 Gesichter nmit einer durchschnittlichen falschen Wiederholungserkennungsrate von unter 10% und einem maximalen Wert von fast 14%. Es gibt somit viele Gesichter mit einer relativ geringen Fehlerrate. Die korrekte Wiederholungserkennungsrate fällt relativ drastisch. Bereits bei den besten 100 Gesichtern beträgt die durchschnittliche

Wiederholungserkennungsrate unter 80% und bei den besten 5.000 Gesichtern hat das schlechteste Gesicht eine Wiederholungserkennungsrate von knapp über 50%. Dies stellt im Alltag somit eine Herausforderung dar. Eine längere Lernphase ist damit zu erwarten bei einer Nutzung.

Bei der Betrachtung der beiden Raten pro Gesicht wird deutlich, dass fast 3.000 Gesichter besser oder gleich dem Durchschnitt in Bezug auf beiden Raten sind. Dies bedeutet, dass deren falsche Wiedererkennung kleiner gleich dem durchschnittlichen Wert ist und die korrekte Wiedererkennung größer gleich dem durchschnittlichen Wert ist. Allerdings nimmt diese Anzahl sehr schnell ab. Beispielsweise gibt es nur 28 Gesichter mit einer falschen Wiederholungserkennungsrate von 5% oder kleiner und einer korrekten Wiedererkennnungsrate von 70% oder größer.

Einteilung der Gesichter

Für jedes Gesicht wurde die korrekte Wiederholungserkennungsrate und die falsche Wiederholungserkennungsrate bestimmt. Mittels der Mittelwerte der beiden Raten erfolgt eine Unterteilung in unterdurchschnittliche bzw. überdurchschnittliche Gesichter und es ergeben sich vier Gruppen:

1. **Werbe**-Gesicht: Diese Gesichter zeichnen sich durch eine überdurchschnittliche korrekte und falsche Wiederholungserkennungsrate aus. Eine Person erkennt dieses Gesicht unabhängig davon, ob dieses vorher gesehen wurde oder nicht.
2. **Geheimnisvolles** Gesicht: Diese Gesichter zeichnen sich durch eine unterdurchschnittliche korrekte und falsche Wiederholungserkennungsrate aus. Eine Person erkennt dieses Gesicht unabhängig davon, ob dieses vorher gesehen wurde oder nicht, tendenziell nicht.
3. **Bekanntes** Gesicht: Diese Gesichter haben eine überdurchschnittliche korrekte Wiederholungserkennungsrate und eine unterdurchschnittliche falsche Wiederholungserkennungsrate. Diese Gesichter werden von einer Person eher erkannt, wenn das Gesicht bereits bekannt ist.
4. **Kurioses** Gesicht: Diese Gesichter haben eine unterdurchschnittliche korrekte Wiederholungserkennungsrate und eine überdurchschnittliche falsche Wiederholungserkennungsrate. Dieses Verhalten ist eher kurios und verleitet zu vielen Fehlern.

Die Einteilung ist angelehnt an eine Bemerkung von Khosla et al. [68].

Abbildung 6.2 stellt diese Werte für jedes Gesicht einzeln dar. Dabei wurde der jeweilige Mittelwert als eine rote Linie eingezeichnet. Links oben sind die kuriosen Gesichter mit einer hohen falschen Wiederholungserkennungsrate und einer geringen Wiederholungserkennungsrate (Anzahl der Gesichter: 2.641). Diese Gesichter sind im Verfahren besonders gefährlich, weil sie einen Angriff erleichtern und gleichzeitig nur sehr schlecht wiedererkannt werden. Sie sind in dem Verfahren zu vermeiden. Rechts oben sind die Werbegesichter mit einer hohen falschen und korrekten Wiederholungserkennungsrate (Anzahl der Gesichter: 1.964). Diese Gesichter werden wiedererkannt, unabhängig, ob sie vorher bereits gesehen wurden oder nicht. Dies ist für Werbung besonders gut, aber erleichtert Angriffe eben so. Unten Links sind

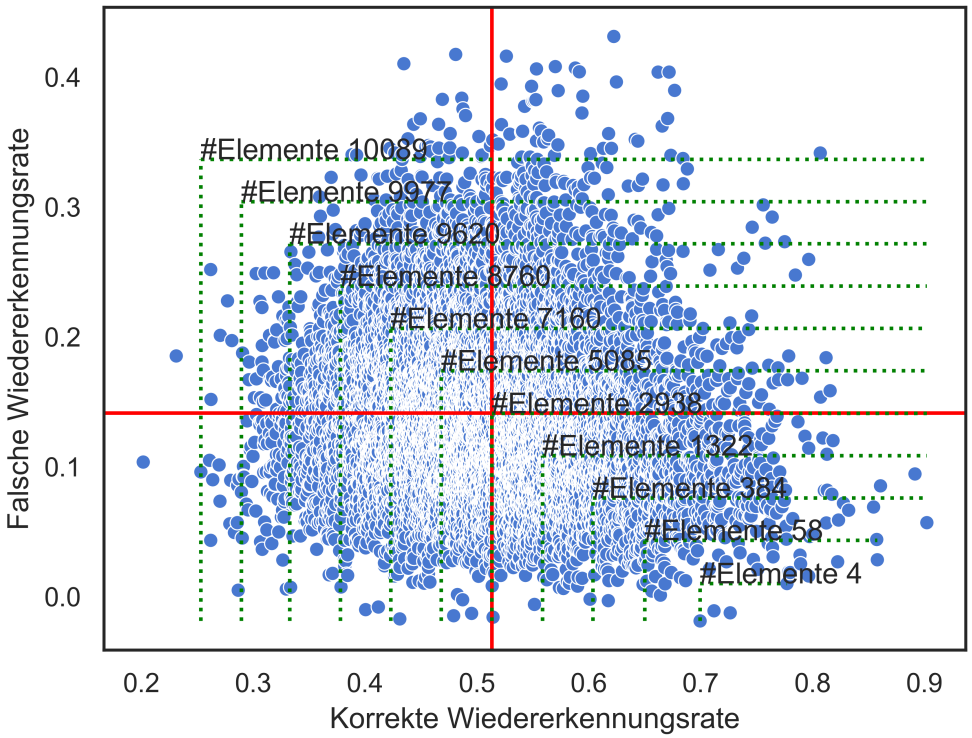

Abbildung 6.2: Vergleich von falscher und korrekter Wiederholungserkennungsrate je Gesicht. Die roten Linien illustrieren den jeweiligen Mittelwert. Besonders geeignet sind die Gesichter im rechten unteren Quadranten. Die grünen Linien illustrieren Teilmengen.

i	Anzahl der Gesichter	Durchschnittliche FAR	HR
3	0		
2.5	0		
2	4	-0.007926	0.728723
1.5	58	0.026879	0.701562
1	384	0.051651	0.669972
0.5	1322	0.073692	0.630296
0	2938	0.091328	0.595001
-0.5	5085	0.106704	0.56534
-1	7160	0.118742	0.540944
-1.5	8760	0.128675	0.524252
-2	9620	0.134621	0.515989
-2.5	9977	0.13807	0.513394
-3	10089	0.139701	0.513065

Tabelle 6.6: Die Auswahl der Gesichter wird die FAR und die HR beachtet und es werden Gesichter ausgewählt, welche einen bestimmten Schwellwert unterschreiten. Die Anzahl ausgewählte Menge an Gesichtern sowie die durchschnittliche falsche Wiederkennungsrate und durchschnittliche korrekte Wiedererkennungsrate der ausgewählten Gesichter werden angezeigt. Dieser Schwellwert wird in Abhängigkeit vom Mittelwert und der Standardabweichung berechnet. Die Berechnung ist: $mean \pm i \cdot std$

die geheimnisvollen Gesichter mit einer geringen falschen und korrekten Wiederholungserkennungsrate (Anzahl der Gesichter: 2.625). Diese verhindern zwar einen Angriff im Durchschnitt besser, aber die korrekte Wiedererkennung unterdurchschnittlich und somit nicht hilfreich bei der Erkennung von legitimen Nachrichten. Unten rechts sind die bekannten Gesichter mit einer geringen falschen Wiederholungserkennungsrate und einer hohen korrekten Wiederholungserkennungsrate (Anzahl der Gesichter: 2.938). Diese Gesichter sind besonders geeignet für das Verfahren, denn Angriffe werden überdurchschnittlich verhindert und die korrekte Wiedererkennung bei legitimen Nachrichten ist überdurchschnittlich gut. Dies sind insgesamt fast ein Drittel aller Gesichter und damit eine geeignete erste Auswahl an Gesichtern. Die durchschnittliche korrekte Wiederholungserkennungsrate ist in dieser Teilmenge $0,6$ (std$= 0,06$) und die durchschnittliche falsche Wiederholungserkennungsrate ist $0,09$ (std$= 0,03$). In Abbildung 6.2 sind mehrere Teilmengen mit der Anzahl der Elemente dargestellt. Dabei wurde ausgehend von dem Mittelwert unter Berücksichtigung der Standardabweichung eine Teilmenge ausgewählt. Die Formel für die Auswahl der Teilmengen war dabei wie folgt:

$$max_{fpr} = mean_{fpr} - i \cdot std_{fpr} \tag{6.1}$$

$$min_{tpr} = mean_{tpr} + i \cdot std_{tpr} \tag{6.2}$$

, wobei $i \in \{-3, -2.5, -2, \ldots, 3\}$. Tabelle 6.6 beschreibt die jeweiligen Teilmengen.

Die Anzahl der Gesichter in den jeweiligen Teilmengen steigt anfangs sehr schnell und zeigt, dass die Anzahl der Gesichter nach Bedarf angepasst werden können und die Wahl einer Teilmenge optimiert werden kann. Die geschätzten korrekten und falschen Wiederholungserkennungsraten und die Wahl von

Teilmengen bildet für die folgende Sicherheitsabschätzung die Grundlage und das Fundament.

6.4.5 Sicherheitsabschätzung

Im Abschnitt 6.3 wurde die Beziehung zwischen dem Sicherheitsspiel Origin und dem Kognitionsspiel $RepeatRecog$, welches dem Experiment von Isola et al. [64] sehr nahe kommt, untersucht. Bei der Kette von Spielen wurde auf das Gedächtnis hingewiesen, Setup fixiert und so genutzt, dass die Angreiferinnen nur zufällig das Verfahren gewinnen kann.

Eine Angreiferin $\mathcal{A}$ gewinnt beim Ergebnis Origin$(1,0)$ im Sicherheitsspiel. Dies entspricht dem Ereignis eines falschen Treffers im Kognitionsspiel $RepeatRecog$ und im Gedächtnis muss der falsche Treffer der legitimen Identität zugeordnet werden. Die Wahrscheinlichkeit eines falschen Treffers kann mit der durchschnittlichen FPR geschätzt werden. Die falsche Zuordnung zu der legitimen Herkunft wird durch eine Gleichverteilung unter den möglichen Elemente abgeschätzt. Die Wahrscheinlichkeit über die Täuschung der Herkunft kann damit für durchschnittliche Nutzerinnen, Angreiferinnen, welche vrf nicht berechnen können, dem fixierte Setup und einer fixierten Anzahl von Gesichtern n, wie folgt geschätzt werden:

$$\beta \geq \Pr[\mathrm{Origin}^{\mathcal{A},\mathcal{H}^{\mathrm{origin}}}_{\mathrm{Setup,Leak,Gen,R}_S} = 1,0 | b = 0] \approx \frac{1}{n} \cdot mean^n_{fpr}$$

,wobei $mean^n_{fpr}$ die durchschnittliche FPR für die n gewählten Elementen ist.

Die Akzeptanz kann auf ähnliche Weise geschätzt werden. Neben den obigen Annahmen wird angenommen, dass die Zuordnung immer korrekt ist. Die Zuordnung bei häufigen Kontakten wird gelernt. Die Wahrscheinlichkeit der Akzeptanz ist wie folgt abgeschätzt:

$$\alpha \leq \Pr[\mathrm{Origin}^{\mathcal{A},\mathcal{H}^{\mathrm{origin}}}_{\mathrm{Setup,Leak,Gen,R}_S} = 1,1 | b = 1] \approx mean^n_{tpr}$$

, wobei $mean^n_{tpr}$ die durchschnittliche FPR für die n gewählten Elementen ist. Die Brauchbarkeit kann dann wie folgt abgeschätzt werden:

$$\epsilon = \frac{\alpha}{\beta} \approx \frac{mean^n_{tpr}}{\frac{1}{n} \cdot mean^n_{fpr}} \geq \frac{mean^n_{tpr}}{mean^n_{fpr}}$$

Die Abschätzung kann sinnvoll sein, weil bei steigenden n dies der dominierende Faktor ist.

Mögliche Darstellungsmengen werden in Tabelle 6.6 dargestellt. Ein pragmatischer Ansatz ist, die Darstellungsmenge iterativ zu vergrößern, und kann durch eine Absenkung von i erfolgen. Die erste Menge kann dann 58 Elemente umfassen und i hat den Wert $1,5$. Danach wird schrittweise i bis $-1,5$ reduziert und danach werden alle Elemente genutzt. Abbildung 6.3 zeigt, dass die Chance für einen erfolgreichen Angriff trotz steigender falscher Erkennungsrate tendenziell geringer wird.

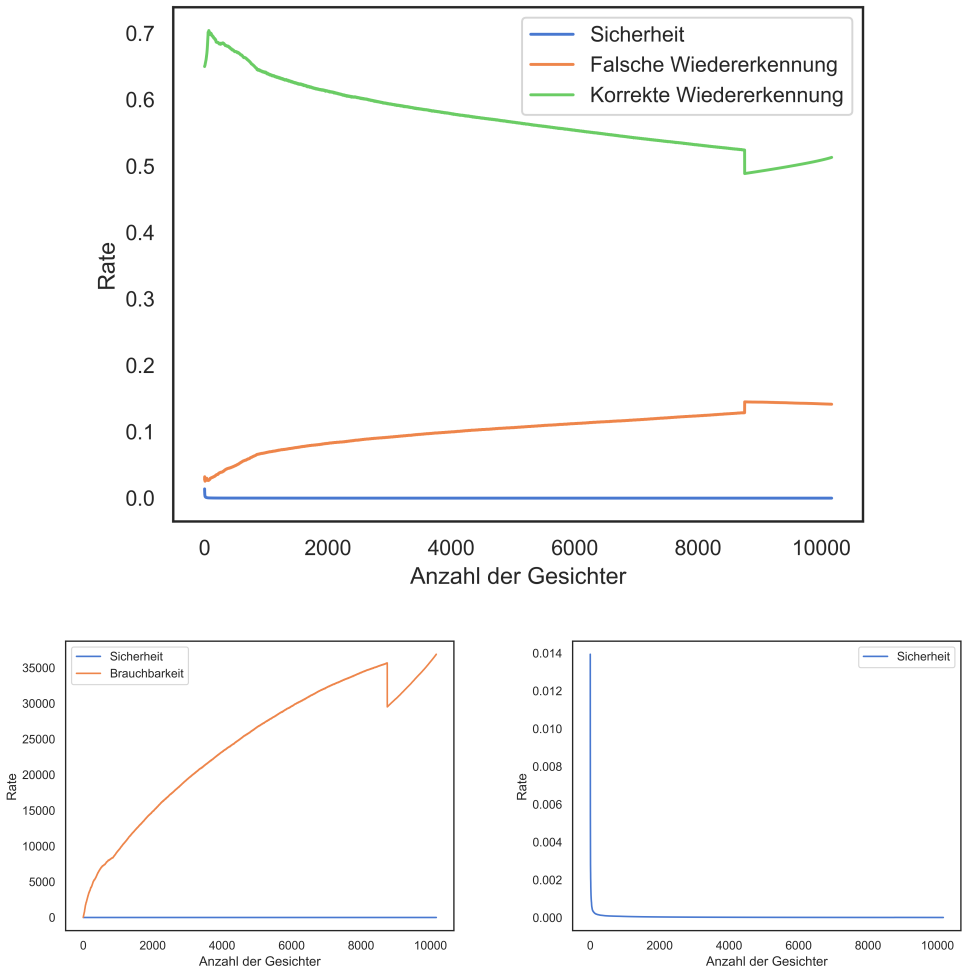

Abbildung 6.3: Die Sicherheitsabschätzung und Brauchbarkeitsabschätzung im Vergleich.

Die Erfolgschance beträgt minimal $1 \cdot 10^{-05}$ und maximal $0,014$, wobei der durchschnittliche Wert $4 \cdot 10^{-05}$ beträgt. Die Wahrscheinlichkeit für einen erfolgreichen Angriff sinkt rapide mit einer steigenden Anzahl an Kontakten.

Die durchschnittliche korrekte Wiederholungserkennungsrate sinkt von $0,704$ auf $0,489$ und ist somit eine deutliche Schwäche von dem Verfahren. Der durchschnittliche Brauchbarkeitswert beträgt $24.317,2$. Mit einer steigenden Anzahl an ausgewählten Gesichtern steigt die Brauchbarkeit. Die Ursache ist der dominierende Faktor $\frac{1}{n}$ durch die zufällige Zuordnungen. Die sinkende korrekte Wiedererkennungsrate bei den gewählten Gesichtern beeinflusst die Schätzung in geringerem Maß. Der Brauchbarkeitswert unterscheidet sich deutlich von 1 und ist damit sinnvoll zur Herleitung einer Herkunft[4]. Die durchschnittliche falsche Wiedererkennungsrate ist immer deutlich kleiner als die durchschnittliche korrekte Wiedererkennungsrate und das Verfahren ist selbst ohne den Faktor $\frac{1}{n}$ brauchbar. In der Praxis ist eine Gefahr, dass die Zuordnung dann durch den Inhalt der Nachricht beeinflusst wird und nicht mehr unabhängig ist. Das dargestellte Verfahren ist somit für durchschnittliche Nutzerinnen unter der Definition von Setup und der Annahme, dass $\mathcal{A}$ nicht vrf überwinden kann, sicher. Die genauen Sicherheitsparameter sind von der Anzahl der Kontakte abhängig. Im Gegensatz dazu ist die einfache Darstellung der E-Mail-Adresse für Angreiferinnen leicht angreifbar und das Verfahren ist unter diesem Aspekt ein Sicherheitsgewinn.

6.4.6 Einschränkungen

Die vorgestellte Methode hat einige inhärente Einschränkungen. Die Daten, mit denen das Modell erstellt wurde, wurden experimentell bestimmt. Das grundlegende Experiment weicht deutlich von dem Sicherheitsszenario ab und berücksichtigt einige Aspekte, wie das Lernen von Darstellungen durch regelmäßige Wiederholungen. Gleichzeitig war dieses Experiment eine Herausforderung für die Teilnehmenden, denn das Experiment überstieg die Möglichkeiten vom Kurzzeitgedächtnis [10] und ist im Allgemeinen nicht der Praxis nahe, weil nur in wenigen Situationen Menschen innerhalb von 4 Minuten mehr als 100 Gesichter betrachten. Eine weitere Einschränkung ist, dass in dem Experiment die Personen keine Identität mit der Darstellung berücksichtigen mussten und somit unklar ist, mit welchem Gesicht die Verwechselung erfolgte.

Durch die steigende Anzahl an Gesichtern kann die Fehlerrate ebenso ansteigen.

Die Sicherheitsabschätzung basiert nur auf der Darstellung der Gesichter und es werden weitere UI-Elemente nicht berücksichtigt. Die Sicherheitsabschätzung lässt somit keine Übertragung in die Praxis zu, aber einen Vergleich mit anderen Verfahren.

Bei der Analyse wurde Setup fixiert und nur dafür wurde die Sicherheit betrachtet. Das Szenario von Setup ist nicht sehr praxisnah, aber eine allgemeine Herausforderung für Menschen.

[4] Dies entspricht einer abgewandelten Interpretation aus dem medizinischen Kontext [41].

Abbildung 6.4: Darstellung der Herkunft in der Mail-App auf iOS.

6.5 Umsetzung

In diesem Abschnitt wurde bisher ein abstraktes Verfahren betrachtet. In diesem Abschnitt wird eine praktische Umsetzung skizziert. E-Mail-Anwendungen, Instant-Messanger und ähnliche Anwendungen stellen neben dem Namen oder der Adresse noch ein rundes UI-Feld für die Herkunft dar. In diesem UI-Feld werden standardmäßig die Initialen dargestellt. Wenn Nutzerinnen zu der Adresse in einem Kontaktbuch ein Foto zu der Person hinterlegt haben, wird dieses angezeigt.

Abbildung 6.4 illustriert dies am Beispiel der Mail-App auf iOS. Die Nutzung des richtigen Gesichts einer Herkunft wird durch die obige Literatur unterstützt und ist die Ideallösung, wenn diese nicht durch die Angreiferin gewählt werden kann. Die Auswahl des Gesichts erfolgt somit in der Regel manuell von den Nutzerinnen und ist damit mit Aufwand für Nutzerinnen verbunden. Es ist somit nicht naheliegend, dass das von der Mehrheit der Nutzerinnen vollumfänglich gemacht wird. Das vorgestellte Verfahren bietet hierzu eine Alternative zur Standarddarstellung. Insbesondere dann, wenn die Person nicht persönlich bekannt ist, kann dies eine sinnvolle Alternative sein. In dieser Umsetzung wird aber neben dem Gesicht noch ein Text angezeigt und dieser kann, wie im Kapitel 5 beschrieben, manipuliert werden. Dies beeinflusst die Sicherheitsbewertung der vollständigen Darstellung. Einerseits erleichtert es das Lernen der zusätzlichen Darstellung und es ist plausibel, dass sehr häufige Darstellungen nur noch an dem Gesicht erkannt werden. Eine falsche Interpretation der E-Mail-Adresse kann Nutzerinnen auffallen, wenn die Interpretation nicht zu der Darstellung passt. Bei einer konkreten Nutzung des Verfahrens ist eine Nutzungsstudie nötig um diesen und mögliche andere Effekte mit weiteren UI-Elementen zu untersuchen.

Eine alternative Möglichkeit ist die puristische Umsetzung von dem Verfahren, indem nur das Gesicht in der Mailbox und in der Detail-Ansicht zur E-Mail zur Darstellung der Herkunft angezeigt wird. Hierbei fehlen weitere Informationen über die Herkunft, wie zum Beispiel die E-Mail-Adresse. Auf einer weiteren Ansicht nur über die Herkunft der Nachricht können diese und weitere Informationen dargestellt werden und bei der Einordnung kann unterstützt werden. Das kann Fehler durch eine flüchtige Interpretation verhindern. In der Praxis kann es bei der Umsetzung verschiedene Varianten geben, wann und wie zusätzliche Informationen zu der obigen Darstellung angezeigt werden. Bei einer konkreten Implementierung werden somit weitere Entscheidungen getroffen, welche die Sicherheit beeinflussen und dies muss untersucht werden.

Die konkrete Implementierung ist dabei aber vom Designkonzept der Anwendung und der restlichen grafischen Oberflächen einer Anwendung abhängig. Aus diesem Grund ist die Implementierung eine Leistung von mehren Expertinnen im Bereich Design und Entwicklung.

Das vorgestellte Verfahren schützt vor Täuschungen auf dem ersten Blick und ermöglicht somit eine deutlich kritischere Auseinandersetzung mit der Herkunft einer Nachricht. Es wird somit Raum und Zeit für eine genauere Betrachtung der unbekannten oder unvertrauten Herkunft geschaffen. Dies kann mit einer weiteren Ansicht genutzt werden. In dieser Ansicht erfolgt die Interpretation der Herkunft, aber dies muss nicht nur auf Basis der (E-Mail)-Adresse erfolgen. Unter der Bedingung, dass die erste Darstellung Nutzerinnen nicht bekannt ist, wird ein weiteres Sicherheitsspiel gestartet, indem in einer anderen Ansicht die Darstellung interpretiert wird. Dies wird nicht weiterverfolgt, aber für zukünftige Arbeiten wird der Gestaltungsraum kurz skizziert.

Ein wesentlicher Aspekt ist die Fragestellung, welche Maßnahmen die Absenderin und der E-Mail-Provider unternommen haben, um eine identische Fälschung zu verhindern. Das umfasst unter anderem die Protokolle SPF, DKIM, DMARC, OpenPGP oder S/MIME. Daneben ist eine offene Frage, ob die Gefahr einer Verwechselung mit einer bekannten Domain oder Adresse (aus dem Kontaktbuch oder der Kommunikationshistorie) besteht. E-Mail-Adressen sind zwar Zeichenketten, aber durch unterschiedliche Zeichensätze sind diese nicht trivial und merkwürdige Adressen können eben so erkannt werden. Eine Adaption zur der Erkennung von verdächtigen Domains im Browser kann erfolgen.[5]

Die Entscheidung der Nutzerinnen wird damit nicht beeinflusst. Neben der Darstellung der Informationen und deren Auswertung können gleichzeitig Handlungsalternativen dargestellt werden, wie zum Beispiel Bestätigung der E-Mail bei der legitimen Herkunft, Erstellung eines Kontaktbucheintrags.

6.6 Zusammenfassung

In diesem Kapitel wurde ein Sicherheitsverfahren gegen Täuschung der Herkunft vorgestellt. Dieses Verfahren basiert auf der zufälligen Zuordnung zwischen einem Gesicht und einer Herkunft und zeigt das Gesicht als Repräsentation der Herkunft an. Die Sicherheit dieses Verfahrens wurde auf empirische Experimente aus der Kognitionswissenschaft hergeleitet und praktische Implementierungsvarianten mit deren Gefahren wurden diskutiert. Das Verfahren ist damit geeignet, in einer konkreten Anwendung in einer Feldstudie weiter evaluiert zu werden.

Das Verfahren hat eine gewisse Ähnlichkeit zum *One-Time-Pad* aus der Kryptographie. In beiden Verfahren wurden die Angreiferinnen kaum eingeschränkt. Gleichzeitig hat es auch Nachteile. In dem hier vorgestellten Verfahren sind die Nachteile die Fehlerraten. In künftigen Arbeiten kann nach Verbesserungen bei der Wahl der Elemente zur Darstellung gestrebt werden. Eine andere Alternative ist mit Hilfe von Einschränkungen der Angreiferinnen oder weiteren Annahmen einen anderen Algorithmus zur Erzeugung einer Darstellung zu entwickeln.

[5] Beispielsweise beschreibt Google die Erkennung für den Chrome-Browser relativ informell hier: https://chromium.googlesource.com/chromium/src/+/main/docs/idn.md

Kapitel 7

Einordnung in den aktuellen Stand der Forschung

7.1 Meta-Forschung

Die menschen-zentrierte IT-Sicherheit ist im Vergleich zu anderen wissenschaftlichen Disziplinen noch relativ jung und interdisziplinär geprägt. Dies führt zu einer breiten Diskussion, wie Sicherheit untersucht werden kann. Krol et al. heben folgende Anforderungen an das Studiendesign hervor [74]:

1. Studienteilnehmende sollten eine primäre Aufgabe haben
2. Studienteilnehmende sollten einer realistischen Gefahr ausgesetzt sein
3. Durchführung von Doppelblindstudien und Unklarheit über das Studienziel
4. Gefahrenmodell, Sicherheit und Benutzbarkeit benötigen genaue Definitionen

Für die Untersuchung eines Verfahrens gegen Phishing-Angriffe sind dies besonders hohe Hürden. Das Gefahrenmodell bei Phishing ist sehr unterschiedlich, denn die Angriffe können gegen eine sehr große Menge an Personen erfolgen oder individuell zugeschnitten sein. Gleichzeitig bietet der dargestellte formale Ansatz eine genaue und präzise Definition der Sicherheit in einem bestimmten Kontext. Bei Studien zur Untersuchung eines Verfahrens und deren Darstellung zur Unterstützung gegen Phishing-Angriffe besteht immer die Gefahr, dass das Ziel der Studie offensichtlich wird. Denn im Gegensatz zum Stimulus in der Medizin ist dieser für die Teilnehmenden sichtbar. Diese Problematik besteht auch bei einer Primäraufgabe sowie der Unklarheit beim Studienziel. Im Kontext von Phishing und insbesondere von Ransomware sind realistische Gefahren innerhalb von Studien schwer umsetzbar und ethisch sehr fragwürdig.

Die Literatur zum Thema Phishing im Allgemeinen sowie mit Bezug zur menschen-zentrierten Sicherheit

ist umfangreich und vielseitig. Zur Aufarbeitung dieser Literatur gibt es zwei aktuelle Veröffentlichungen zur Systematisierung der Phishing-Literatur mit einem Fokus auf den Menschen [47, 151]. Franz et al. haben durch eine Literaturrecherche mehr als 2.000 Veröffentlichungen gefunden, aus denen sie 64 mit menschen-zentrierten Interventionen gegen Phishing identifiziert haben [47]. Zhuo et al. haben aus mehr als 4.000 Veröffentlichungen 45 mit einem menschen-zentrierten Ansatz ausgesucht [151]. In der Metastudie von Franz et al. waren die betrachteten Studien häufig Laborstudien (20 mal) oder Feldstudien (16 mal). Zur Untersuchung von neuen Verfahren ist dies eine besondere Herausforderung. Denn das Verfahren muss zunächst erlernt werden und besondere Darstellungen bei Angriffen kann das Verhalten verändern, obwohl in der Praxis die besondere Darstellung häufiger ein falsch positiver Fall ist [147]. Franz et al. heben insbesondere hervor, dass nur ein Drittel der Studien unter realistischen Bedingungen erfolgte [47]. Die Schaffung von realistischen Bedingungen sind aus den obigen Gründen wiederum eine Hürde bei der Untersuchung von einem neuen Verfahren. Insbesondere die regelmäßige Nutzung ist ein wichtiger Aspekt. Zusätzlich nahmen an den Feldstudien häufig Personen aus dem universitären Kontext teil [47]. In der Meta-Studie von Franz et al. ist der zweithäufigste betrachtete Angriffsvektor die E-Mail (17 mal) und zwar nach der URL (33 mal), aber deutlich vor der Webseite (zehnmal) und Authentifizierung (zwölfmal) [47]. Dies verdeutlicht das Forschungsinteresse an der E-Mail als ein Angriffsvektor. Der Untersuchungsgegenstand war dabei meist ein Sicherheitstraining (14 mal von insgesamt 30 Publikationen zum Sicherheitstraining), Schulung (viermal von insgesamt 7 Publikationen zu Schulungen), Schaffung von Achtsamkeit (dreimal von insgesamt 17 Veröffentlichungen) und UI-Design-Vorschlägen (zweimal von insgesamt 20 Veröffentlichungen mit UI-Design-Vorschlägen) [47]. Bei einer der beiden Veröffentlichungen zum Design ist der Autor dieser Dissertation der Erstautor. Bei Sicherheitstraining und Schulungen sind die meisten Veröffentlichungen zum Thema E-Mail, wohingegen bei konkreten Verbesserungsvorschlägen vom UI oder Warnungen sowie Erhöhung der Achtsamkeit die E-Mail unterrepräsentiert ist. Dabei sind, wie an dem vorgestellten Verfahren deutlich wird, viele Verbesserungsvorschläge am Design der E-Mail möglich. Ein Erklärungsansatz für dieses Ungleichgewicht ist die fehlende Möglichkeit, Verfahren effektiv, nachvollziehbar und glaubwürdig zu untersuchen. Eine Kernaussage von Zhou et al. [151] lautet:

> Current anti-phishing training is not effective in protecting users from phishing attacks. It is essential for future research to investigate approaches that truly help the users reduce phishing susceptibility.

Dies zeigt in Kombination mit dem aktuellen Ungleichgewicht bei den Untersuchungsgegenständen nochmals den Bedarf an Forschungsmethoden und Ansätzen, um neue Verfahren zur Erkennung von Phishing zu entwickeln.

Zhou et al. formulieren folgende Hoffnung und Wunsch an weitere Forschungen:

> We see a great opportunity for research to adopt this technology in understanding user behaviour and using psychological theories to reason about the findings. Regarding phishing detection, we found several areas that have been studied in psychology, and can be applied to human-centred phishing studies to help to explain the variations in users' phishing suscepti-

> bility. As such, we hope future studies can focus on users' mental processes to understand the variables that can influence their decision and impact their phishing detection performance. From the reviewed studies, we also found a lack of discussion on tools or systems that aim to help users determine email legitimacy when phishing emails arrived in their mailbox.

Diese Arbeit hat diesen Aspekt aufgenommen und bietet einen Vorschlag für eine Forschungsmethode sowie ein erstes Verfahren. Kaur et al. haben im Allgemeinen beobachtet, dass menschen-zentrierte Sicherheitsforschung meist explorativ erfolgt und kein konstruktiver Ansatz verfolgt wird [67].

7.2 Warnungen

In der menschen-zentrierten Sicherheit ist eine Warnung der Nutzerinnen eine Standardmaßnahme und wird häufig angewendet. Insbesondere im Kontext von fehlender Transportverschlüsselung wurden Warnungen im Browser häufig untersucht [3, 108, 129]. Beispielsweise konnte Google die Beachtung der Warnungen steigern, indem Warnungen bei einer Webseite seltener angezeigt wurden [142]. Bei einer fehlenden Transportverschlüsselung wurden Nutzerinnen mittels Warnungen informiert und es gibt klare Anweisungen zur Vermeidung der Gefahr. Zum Umgehen der Warnung und dem Besuch der unverschlüsselten Webseite sind deutlich mehr Interaktionen der Nutzerinnen nötig oder es ist ganz unmöglich.

Die Nutzerinnen können abwägen, ob die Webseite trotzdem besucht werden soll oder nicht. Diese Abwägung kann eine Einzelfallentscheidung für jede Webseite sein. Der Besuch einer Informationsseite, wie zum Beispiel Restaurantkritiken oder ein Theaterprogramm, ist unverschlüsselt für die Mehrheit von Nutzerinnen sinnvoll. Denn selbst wenn dieser Besuch abgehört wird, beeinträchtigt es normale Internetnutzerinnen nicht und ist ungefährlich. Die Eingabe von Passwörtern oder anderen Informationen hingegen ist selbst für normale Internetnutzerinnen auf einer unverschlüsselten Webseite potentiell gefährlich. Bei einer Warnung zur fehlenden Transportverschlüsselung können Nutzerinnen abwägen. Sowohl der Besuch oder das Fernbleiben von der Webseite ist je nach Situation sinnvoll. Insbesondere unter der Annahme, dass Webseiten langfristig einen bestimmten Dienst anbieten, ist es sinnvoll, die Abwägung der Nutzerinnen für einige Zeit beizubehalten und nicht jedesmal eine neue Abwägung zu fordern. Im Hinblick auf eine Verbesserung einer Browserwarnung kann es sogar sinnvoll sein, Sicherheitspolitiken im Browser einzuführen. Vor der Eingabe von Daten auf unverschlüsselten Webseiten können Nutzerinnen gewarnt werden oder die Eingabe ist gar nicht möglich. In Anbetracht der Verbreitung von Transportverschlüsselung im Web ist dies aber kaum nötig.[1]

Im Gegensatz zu einer fehlenden Transportverschlüsselung ist die Erkennung von einem Phishing-Angriff nicht fehlerfrei möglich und ist mit einer gewissen Restunsicherheit verbunden. Falls ein Phishing-Angriff (fehlerfrei) erkannt wurde, gibt es nur eine allgemeine Handlungsanweisung: Die Löschung der Nachricht oder das Blockieren der Webseite. Die Nutzerinnen müssen nicht mehr mit der Entscheidung konfron-

[1] Beispielsweise meldet Google für den Chrome-Browser teilweise Verschlüsselungsraten von 99%. https://transparencyreport.google.com/https/overview?hl=en), Letzter Zugriff: 3. Mai 2023, 17:30

tiert werden. Phishing-Nachrichten werden vom Provider direkt abgewiesen werden oder im Spamordner abgelegt. In diesen Fällen gibt es selten eine Warnung, aber nicht alle Angriffe werden vom Provider entdeckt.

Die Restunsicherheit bedeutet, dass entweder legitime Nachrichten als Phishing-Angriffe gewertet werden oder Phishing-Angriffe als legitime Nachrichten eingestuft werden. In einem System, in dem Google innerhalb von wenigen Monaten mehr als 600 Millionen eingehende E-Mails als Phishing-Angriffe identifiziert [123], sind beide Fehlerarten mit hohen Kosten verbunden. Eine Phishing-Warnung erfolgt somit häufig in den nicht ganz eindeutigen Fällen und erfordert eine Entscheidung (legitim oder nicht) von der Nutzerin. Falls eine Nachricht ein Phishing-Angriff ist, ist die allgemeine Handlungsempfehlung keine Links, Anhänge zu öffnen und die Nachricht zu löschen. Dies ist ein wesentlicher Unterschied gegenüber Warnungen vor fehlender Transportverschlüsselung. Die Nutzerinnen müssen entscheiden, ob die E-Mail legitim (also sicher ist) und nicht abwägen, wie sie mit der unsicheren (also unverschlüsselten) Webseite umgehen. Die Nutzerinnen müssen in diesem Fall die Entscheidung vom (Experten)-System validieren.

Die Anzahl an falschen Warnungen muss dabei möglichst gering sein. Ansonsten besteht die Gefahr vor einem Automatismus und dem Ignorieren von Phishing-Warnungen. Gleichzeitig ist bereits ein erfolgreicher Phishing-Angriff (vor dem wahrscheinlich nicht gewarnt wurde) für einen massiven Schaden ausreichend. Phishing-Warnungen befinden sich somit im klassischen Dilemma von falsch positiven Fehlern und falsch negativen Fehlern. Phishing-Warnungen sind damit in ihrem Nutzen deutlich eingeschränkt. Phishing-Warnungen sind üblich in der wissenschaftlichen Untersuchung von Phishing [47] und es gibt viele Beispiele [43, 99, 24]. Das vorgeschlagene Verfahren bietet dafür eine Alternative und die vorgeschlagene Formalisierung zeigt nochmal das Zusammenwirken von Sicherheitsprotokollen und der Darstellung, sowie die Bedeutung von den falsch positiven und falsch negativen Fehlern.

7.3 Verwandte formale Methoden

In der Kryptographie sind formale Methoden akzeptiert und verbreitet. Aus diesem Forschungszweig gibt es selten Überschneidungen mit menschen-zentrierter Sicherheit. Hopper und Blum haben ein Authentifizierungsverfahren zwischen Menschen und Computer auf Basis von einem NP-schweren Problem vorgeschlagen [61]. Ihre Sicherheitsdefinition ist an klassischen Sicherheitsdefinitionen der Kryptographie angelehnt.

Bei einem Authentifizierungsverfahren tätigt der Mensch eine Eingabe und der Computer validiert diese Eingabe. Wie im Verfahren von Hopper und Blum kann es möglich sein, dass der Computer dem Menschen eine Aufgabe stellt und diese mit einem ihm bekannten Geheimnis lösen kann. Die Sicherheit des Verfahrens basiert auf einen Reduktionsbeweis mit einem NP-schweren Problem. Es wird angenommen, dass der Mensch die Berechnungen tätigen kann, aber der Mensch selbst ist nicht Teil des Sicherheitsbeweises. Im Sicherheitsbeweis wird nur der Austausch der Nachrichten zwischen Mensch und Computer berücksichtigt. Stattdessen werden die menschlichen Fähigkeiten bei der Benutzbarkeit betrachtet und

dies wird in Definition 2 wie folgt ergänzt [61]:

> An identification protocol (H,C) is said to be (α, β, t)-human executable if at least a $(1 - \alpha)$ portion of the human population can perform the computations H unaided and without errors in at most t seconds, with probability greater than $1 - \beta$.

Hopper und Blum sehen ihre Definition kritisch, aber benutzen diese in Ermangelung von Alternativen und verweisen auf empirische Experimente. Ein Problem ist hierbei sicherlich festzustellen, welche Fähigkeiten ein bestimmter Prozentsatz der Menschheit hat. Ihr Experiment führten sie unter Studentinnen und Mitarbeiterinnen an ihrem Mathematik-Fachbereich an der Carnegie Mellon University durch. In diesem Experiment konnten die Teilnehmerinnen mit einer erfolgreichen Authentifizierung eine Gratis-Cola erhalten. An ihrer Studien nahmen 54 Personen teil und in 155 von 195 Versuchen waren diese erfolgreich. Für eine durchschnittliche erfolgreiche Authentifizierung benötigen die Personen 166 Sekunden. Sie schlussfolgern daraus [61]:

> Thus it is empirically clear that there is some value α for which this is a $(\alpha, .25, 160)$-human executable identification protocol which is secure against computationally bounded eavesdropping adversaries.

Die Hauptaufgabe in dem Experiment bestand aus komplexen Kopfrechnungen und die Teilnehmerinnen waren Studentinnen oder Mitarbeiter am Mathematik-Fachbereich. Darum ist vermutlich das gesuchte α sehr klein und eine Authentifizierungszeit von mehr als zwei Minuten nicht alltagstauglich für viele Anwendungszwecke, wie zum Beispiel eine Authentifizierung an einem Smartphone, Tablet oder Laptop. Die Implementierung für das Experiment ist nicht bekannt und es ist offen, ob mit einer Unterstützung der Nutzerinnen die Eingabezeit und der Erfolg verringert werden kann und welchen Einfluss dies auf die Sicherheit hat.

Das Verfahren und die Analyse sind sehr stark von der kryptographischen Perspektive geprägt und der Mensch wird nur am Rand betrachtet. Dies hat den Vorteil, dass die Methode und das Verfahren aus der kryptographischen Perspektive nachvollziehbar ist, aber der praktische Nutzen von diesem Verfahren ist dann doch fraglich. Im Gegensatz zu Verfahren, welche aus der menschen-zentrierten Forschung vorgestellt werden, verdeutlicht dieses Verfahren, dass bei Mensch-Computer-Authentifizierungsverfahren ein Spannungsfeld zwischen (beweisbarer) Sicherheit und Benutzbarkeit herrscht und unterschiedliche Schwerpunkte möglich sind [148].

Bei der Erkennung von Phishing-Nachrichten muss dies nicht zwangsläufig sein, weil der Mensch die Entscheidung trifft. Bei dieser Entscheidung sollten die Angriffe erkannt werden und damit dies in der Praxis erfolgen kann, muss es benutzbar sein. Denn eine schwierige Benutzbarkeit erhöht nicht die Sicherheit, wenn der Mensch stattdessen einfach alle Nachrichten als legitim eingestuft.

Im Gegensatz zu der Mensch-Computer-Authentifizierung muss bei Phishing-Angriffen der Mensch bei der Sicherheitsanalyse betrachtet werden. Denn dieser trifft die Entscheidung. In der vorgeschlagenen Analysemethode wird dies berücksichtigt. Damit kann ein Verfahren nicht mehr nur auf kryptographi-

schen Verfahren basieren, sondern muss den Menschen mit dessen Fähigkeiten berücksichtigen. Das vorgeschlagene Verfahren wurde ausgehend von menschlichen Fähigkeiten konstruiert und ermöglicht damit Ergebnisse aus der Kognitionswissenschaften zu verwenden, um das Verfahren zu analysieren. Der Bezug auf die Kognitionswissenschaft und die Fähigkeiten der Menschen ist deutlich ausgeprägter als in der Publikation von Hopper und Blum.

Ebenso wie das Verfahren von Hopper und Blum ermöglicht es zunächst eine abstrakte Analyse von dem Verfahren und hebt die wesentlichen Eigenschaften zur Sicherheit des Verfahrens hervor. In beiden Fällen können die Verfahren adaptiert und implementiert werden[2].

Gajek et al. haben aufbauend auf TLS ein kryptographisches Protokoll zur beidseitigen Authentifizierung von Nutzerinnen und Server entwickelt [48]. In ihrem formalen Modell zum Sicherheitsbeweis modellieren sie den Menschen. Sie führen einen *Human Perceptible Authenticator (HPA)* ein und nehmen an, dass ein Mensch solch einen HPA einfach wiedererkennen kann. Sie definieren informell die Unverwechselbarkeit von zwei HPAs wie folgt [48]:

> We call two authenticators w and $w^\star$ indistinguishable if there is no human user $\mathcal{H}$ that distinguishes (recognizes) with non-negligible probability between the case that a message $m^\star$ contains w and the case that $m^\star$ contains $w^\star$ (denoted in the following by $m^\star|w$ and $m^\star|w^\star$, respectively)

Zur Illustration verweisen sie auf die Forschungsergebnisse zu grafischen Passwörtern und empfehlen die Verwendung von analogen Grafiken. Es fehlt hierbei eine genauere Untersuchung der Annahme und es wird auf die Notwendigkeit zur Durchführung von Experimenten verwiesen.

Beispielsweise hat Valentine das Authentifizierungsverfahren Passface untersucht und in unterschiedlichen Studienbedingungen konnten sich 77% bis 100% der Teilnehmerinnen nach drei Versuchen erfolgreich anmelden [16, 133]. Es ist damit fraglich, ob es überhaupt solche HPAs gibt. Die experimentelle und empirische Bestimmung sowie die Einschränkungen der menschlichen Kognition lassen eine asymptotische Betrachtung (also mit vernachlässigbarer Wahrscheinlichkeit) nur schwer zu. Aus diesem Grund wurde im hier vorgestellten Modell und Verfahren die konkrete und nicht-asymptotische Sicherheit betrachtet.

Gajek et al. nehmen an, dass der Mensch sich ein Passwort merken und ein HPA wiedererkennen kann. Im Anschluss wird ein kryptographisches Protokoll zur gegenseitigen Authentifizierung ohne eine vertrauenswürdige dritte Partei vorgestellt und der Beweis beruht unter anderem auf den obigen Annahmen [48]. Sie fokussieren sich auf die Konstruktion von einem kryptographischen Protokoll und ein wesentlicher Vorteil ist, dass sie die Anforderungen an den Menschen klar benennen. Allerdings vermeiden sie eine detaillierte Auseinandersetzung mit den Fähigkeiten des Menschen, indem sie beispielsweise keine Literatur aus der Kognitionswissenschaft oder Psychologie betrachten. Als einen Vorteil in ihrem Protokoll wird erwähnt, dass Nutzerinnen nicht mehr die URL oder andere Sicherheitsindikatoren betrachten

[2] Erstaunlicherweise wurden die Publikation von Hopper und Blum vor allem von Veröffentlichungen über Authentifizierungsverfahren mit beschränkter Hardware zitiert.

müssen. Eine spannende praktische Fragestellung ist hierbei aber, wie Nutzerinnen (vermeintliche) Konflikte bei der Darstellung lösen. Die Konzepte in dieser Veröffentlichungen wurden im Anschluss nicht mehr weiterverfolgt.

Basin et al. ergänzen die formale Beweisführung von Sicherheitsprotokollen um menschliche Fehler [11]. Dabei werden Menschen als regelbasierte (menschliche) Agenten definiert, welche besonderen Handlungsempfehlungen folgen. Ein Beispiel für solche einfachen Empfehlungen ist, dass private Schlüssel geheim bleiben soll. Es wird angenommen, dass Menschen diesen Empfehlungen folgen. Diese Umsetzung berücksichtigt dabei nicht die menschliche Kognition und Verhalten.

Benenson et al. stellen ein allgemeines formales Modell zur Betrachtung der Sicherheit in der menschenzentrierten Sicherheit vor [13]. Ihr Modell ist sehr formalistisch und allgemein gehalten. Es kann auf verschiedene Problemstellungen angewendet werden. In ihrer Veröffentlichung wird dieses Modell aber nicht an einem Verfahren angewendet, sondern nur sehr allgemein an einem Anwendungsfall (die Authentifizierung mittels zwei Faktoren). Bei der Betrachtung der Publikationen, welche die Veröffentlichungen zitieren, wird dieses Modell nicht zur Analyse eines Verfahrens verwendet. Im Gegensatz dazu ist das vorgeschlagene formale Modell zwar eingeschränkt auf Phishing-Angriffe, aber ermöglicht die Konstruktion und Evaluierung von Sicherheitsverfahren. Gleichzeitig kann das Modell auf andere Probleme adaptiert werden. Benenson et al. favorisieren zur Untersuchung der menschlichen Fähigkeiten ein ECG (Elektrokardiogramm), also der Herzfrequenz. Das in Kapitel 4 vorgeschlagene Modell und die Analyse basieren hingegen auf Erkenntnissen aus der Kognitionswissenschaft. [13] Ein Vorteil dabei ist, dass diese Ergebnisse bereits validiert wurden und zunächst keine eigenen Experimente nötig sind, welche nur selten reproduziert werden. Neue Erkenntnisse in der Kognitionswissenschaft können hingegen sofort einen Einfluss auf den Ausgang der Analyse haben.

Beckert und Beuster ergänzen die formale Verifikation von Programmcodes um eine Variante der GOMS-Methode[3] [12]. Diese Analyse ist sehr formal, komplex und abhängig von der konkreten Implementierung. Gleichzeitig ist die formale Verifikation von Programmcodes die Ausnahme und insbesondere die formale Verifikation von E-Mail-Anwendungen und der verwendeten grafischen Oberflächen-Frameworks erscheinen kaum umsetzbar. Beckert und Beuster zeigen damit einen verwundbaren Aspekt und bieten gleichzeitig einen konstruktiven Lösungsansatz, aber aus den obigen Gründen wurde dieser nicht verfolgt.

Cranor hat ein Modell aus der psychologischen Perspektive entworfen und stellt die kognitiven Abläufe vom Menschen in den Vordergrund [34]. Ihr Modell ist nur schwer mit formalen Modellen aus der Kryptographie zu verknüpfen. Es ist sehr allgemein und an kein konkretes Gefahrenmodell angeknüpft. Aus diesem Grund ist eine Analyse von Verfahren nicht möglich. Es verdeutlicht aber die Einflüsse von kognitiven Prozessen, welche bei der Konstruktion eines Verfahrens berücksichtigt werden müssen und die Sicherheit beeinflussen.

Ellison beschreibt die allgemeine Herausforderung, dass Menschen im Anschluss an ein Netzwerkpro-

[3] GOMS ist eine formale Untersuchungsmethode der Benutzbarkeit von einer grafischen Oberfläche. In dieser werden die Nutzerinteraktionen, um eine bestimmte Aufgabe zu erreichen, gezählt. Weitere Informationen werden von Card et al. beschrieben [28].

tokoll über die Herkunft überzeugt werden müssen und in diesem Schritt Fehler verhindert werden müssen [44]. Er bezeichnet das Netzwerkprotokoll inklusive dieser Erweiterung als *Ceremony* und empfiehlt diese ähnlich wie Sicherheitsprotokolle zu untersuchen. TLS und verschlüsselte bzw. signierte E-Mails werden als Beispiel genannt und die möglichen Probleme werden diskutiert, aber es werden keine (formalisierten) Lösungen präsentiert. Abseits von Phishing betrachten Bonneau und Schechter die Merkfähigkeit von Menschen im Kontext von Passwörtern [20]. Sie nutzen dabei psychologische Erkenntnisse, um die Ergebnisse ihrer Studie zu erklären. Pfleeger und Caputo betonen, dass das menschliche Verhalten berücksichtigt werden muss und illustrieren mittels der psychologischen Erkenntnis, dass Erkennen einfacher als Erinnern ist und dies bei Authentifizierungsverfahren genutzt werden kann.

7.4 Ähnliche Verfahren

Am nächsten zu dem hier vorgeschlagenen Verfahren sind Anti-Phishing-Darstellungen und entsprechende Toolbars im Webbrowser. Als weitere grafische Darstellungen bei der Eingabe von Passwörtern oder Kreditkarteninformationen wurde für den Webbrowser vorgeschlagen.

Dhamija und Tygar schlagen vor, personalisierte Bilder als Hintergrund bei der Eingabe von einem Passwort zu verwenden [39]. Dabei müssen Nutzerinnen jeweils ein individuelles Bild merken. Das jeweilige Bild wird für alle Webseiten verwendet und die Online-Dienste müssen das System unterstützen. In einem weiteren Verfahren berechnen Server und Browser jeweils ein Bild und der Mensch kann entscheiden, ob sie gleich sind. Wenn Nutzerinnen ungleiche Bilder erkennen, findet sehr wahrscheinlich ein Angriff statt.

Adelsbach et al. schlagen personalisierte Bilder für die Darstellung von Sicherheitsindikatoren im Browser vor [1]. Damit wird verhindert, dass ein Angreifer Sicherheitsindikatoren nachahmen kann.

Diese Verfahren fokussieren sich auf den Webbrowser und die Eingabe von Passwörtern. In deren Darstellungen sind die Verfahren eine Ergänzung zur URL. Die Herleitung der Herkunft der Webseite ist nicht über die Indikatoren möglich. Diese unterstützen bei der Einstufung der Sicherheit. Das hier vorgeschlagenen Verfahren hingegen unterstützt Nutzerinnen bei der Herleitung der Herkunft.

Die Browser-Toolbar von Herzberg und Jbara visualisiert die Logos von Unternehmen, welche mittels einer PKI ermittelt wurden [59]. Ihr Vorschlag beruht damit im Gegensatz zu dem hier vorgeschlagenen auf einer vertrauenswürdigen dritten Partei.

Kapitel 8

Schluss

Die vorherigen Kapitel zeigten die Bedeutung von Phishing-Angriffen in der Praxis und in der Forschung. Seit mehr als zwei Jahrzehnten ist es eine dauerhafte Erscheinung im Internet, die sich mit der Zeit verändert, aber nicht an Relevanz verloren hat. Bei den aktuellen Forschungsarbeiten wird häufig die Erkennung der Symptome durch Menschen von einem Phishing-Angriff thematisiert. Dabei wird oftmals nicht beleuchtet, wie Phishing-Angriffe im Kern verhindert werden können und Darstellungen und Anwendungen abseits von Warnungen gestaltet werden können.

Im Gegensatz zu den bisherigen Forschungen aus der menschen-zentrierten Sicherheit hat diese Arbeit mit der Formalisierung und Modellierung von Phishing-Angriffen begonnen. Dadurch wird die Perspektive auf das Wesentliche bei einem Phishing-Angriff verschoben:

1. Der Erfolg eines Angriffes ist abhängig von der Entscheidung der Nutzerinnen
2. Die Einbeziehung der vergangenen Kommunikation zwischen den Nutzerinnen und den legitimen Parteien
3. Die Steigerung der Erfolgswahrscheinlichkeit für einen Angriff durch das Ausnutzen von kognitiven und psychologischen Schwächen

Die Berücksichtigung der menschliche Entscheidung ist die größte Herausforderung bei der Modellierung und Formalisierung für die Sicherheit. Dies zeigt, dass zwar Maßnahmen durch Service-Provider die Gefahren reduzieren können, aber schlussendlich muss die grafische Oberfläche die Nutzerinnen bei der Entscheidung unterstützen und die Nutzerinnen müssen die Gefahren erkennen können. Die Vermittlung von Wissen sowie die Sensibilisierung der Nutzerinnen ist die gängige Empfehlung aus der menschenzentrierten Sicherheitsforschung. Die alltägliche Erfahrung mit E-Mail-Anwendungen verdeutlicht, dass Nutzerinnen nur selten bei der Erkennung von Angriffen unterstützen werden. Die gemeinsame unveröffentlichte Arbeit mit dem KIT bestätigen dies. Die Verbesserungen der grafischen Darstellungen einen Nachricht bieten weiteres Potential zur Unterstützung der Nutzerinnen bei der Bearbeitung ihrer

Nachrichten und können die Sicherheit erhöhen.

Die Bekanntheit der legitimen Partei, welche beim Angriff imitiert wird, ist essentiell bei Phishing-Angriffen und hebt diesen von anderen unerwünschten Nachrichten, wie zum Beispiel Spam, ab. Zur Vereinfachung wird angenommen, dass bereits in der Vergangenheit eine Kommunikation zwischen Nutzerinnen und den legitimen Parteien erfolgte. Es verdeutlicht die Herausforderung bei der Durchführung von Phishing-Experimenten mit einer neuen Anwendung. Die allgemeinen Erkennung von Angriffen durch den Service-Provider (zum Beispiel die Blockliste von Google oder ML-Techniken zur Erkennung von Phishing) kann nicht immer die vergangene Kommunikation berücksichtigen und bewerten. Die gemeinsame Vergangenheit als ein wesentlicher Aspekt vom Kontext, in dem der Angriff stattfindet, ermöglicht Endanwendungen, Nutzerinnen zu unterstützen. Dafür ist es erforderlich, dass die Endanwendung über einen Zustand über die vergangene Kommunikation verfügt. Im Alltag haben Nutzerinnen ebenso eine Art Zustand über die vergangene Kommunikation. In Experimenten und in der Formalisierung muss dies berücksichtigt werden. In der Formalisierung wurde diese mit der Funktion Setup umgesetzt und zeigt gleichzeitig die Herausforderung bei der Durchführung von Experimenten.

Die Formalisierung in Kapitel 4 zeigt die Bedeutung der kognitiven Fähigkeiten und Einschränkungen von Menschen, welche eben die Nutzerinnen sind, und die Konsequenzen für die Sicherheit. Dabei wird hervorgehoben, dass bereits bei der Entwicklung von Darstellungsverfahren dies ohne die Durchführung von (aufwendigen und zeitintensiven) Experimenten mit Nutzerinnen berücksichtigt werden kann. Die hier vorgeschlagene Modellierung bietet eine Schnittstelle zu den allgemeinen Erkenntnissen und Experimenten aus der Kognitionswissenschaft und Psychologie.

Die Formalisierungen kombinieren grundlegende Ansätze aus der modernen (theoretischen) Kryptographie mit Erkenntnissen aus Experimenten zur Kognition und der Untersuchung von (medizinischen) diagnostischen Verfahren. Bereits durch das Aufzeigen dieser Verbindungen hat die Formalisierung einen Mehrwert. Sie bietet eine neue Perspektive und Herangehensweise an das Phishing-Problem. Damit werden neue Lösungen und neue Analysemethoden aufgezeigt. Sie ergänzt damit die bestehenden Methoden und Perspektiven der Forschungsgemeinschaft.

Die vorgeschlagene Formalisierung wurde auf zwei Arten angewendet. Einerseits wurde allgemeine Schwächen von Darstellungen von E-Mail-Adressen hervorgehoben. Andererseits wurde ein konkretes Verfahren für eine bessere Darstellung entwickelt und untersucht. Die Formalisierung und die entwickelten Heuristiken bieten eine allgemeine Betrachtung der Problematik und stellen die menschliche Wahrnehmung in den Vordergrund. Die zugrunde liegende kognitive Verarbeitung, welche in den verschiedenen Spielen und Experimenten zu Tage treten, wurde in der bisherigen Forschung vernachlässigt. Die Ausdifferenzierung der verschiedenen kognitiven Verarbeitungsebenen ermöglicht die Fokussierung und Untersuchung einzelner kognitiver Fallstricke und mögliche Gegenmaßnahmen. Die allgemeinen Prinzipien, welche sich hinter der technischen Umsetzung verbergen, wurden in der bisherigen Forschung vernachlässigt. In der Vergangenheit wurden die Gefahren immer und immer wieder in Anwendungen wiederentdeckt. Die Heuristiken bieten hingegen bereits bei der Entwicklung eines Verfahrens die Aufdeckung von möglichen Schwächen. Die Verallgemeinerung hilft somit bei künftigen Verfahren.

Exemplarisch wurde die E-Mail-Adresse betrachtet. Die Darstellung von E-Mail-Adressen wurde in Beziehung zu den bekannten Schwächen der menschlichen Wahrnehmung und Verarbeitung von Zeichenketten gesetzt. In einem historischen Datensatz von Phishing-Angriffen wurden Angriffe identifiziert, welche diese Schwächen ausnutzen wollten. Die Erkenntnisse aus der Kognitionswissenschaft und der Heuristik hätten bereits bei der Entwicklung der Protokolle und der Darstellungen helfen können, diese Gefahren zu erkennen und nach Möglichkeit zu berücksichtigen. Beispielsweise haben Apple und Mozilla erst in den letzten Jahren Gegenmaßnahmen zur Täuschung mittels des Anzeigenamen bei der E-Mail-Adresse umgesetzt.

Das Kapitel 5 verdeutlicht die Gefahren bei der Verarbeitung von dargestellten E-Mail-Adressen durch die Berücksichtigung der kognitiven Forschung zur Verarbeitung von Zeichen und Zeichenketten. Künftig sollten die Heuristiken bei der Entwicklung von Protokollen und Anwendungen besser berücksichtigt werden.

Im Kapitel 6 wird die Formalisierung konstruktiv verwendet, um ein Darstellungsverfahren zu modellieren. Es wird vorgeschlagen, die Darstellung der Herkunft einer Nachricht um ein zufälliges Gesicht zu ergänzen oder gar zu ersetzen, um die Wiedererkennung zu verbessern. Die Darstellung wurde aufgrund von kognitiven Eigenschaften ausgewählt. Die Sicherheit wurde durch allgemeine und etablierte kognitive Experimente und Theorien gezeigt. Durch die exakte formale Beschreibung werden die notwendigen Prinzipien für die Anwendung entwickelt. Diese sind:

1. Die Darstellung sollte nicht durch eine unbekannte dritte Partei (zum Beispiel der Absenderin) beeinflussbar sein.
2. Die Darstellung sollte leicht von Nutzerinnen wiedererkannt werden.
3. Die Darstellung sollte keine komplexe Semantik haben.

Das vorgestellte Verfahren nutzt diese Prinzipien. Zur der Untersuchung der Sicherheit sind diese explizit nötig. Eine konkrete Implementierung der Darstellung muss dies berücksichtigen, wobei Abwägungen möglich sind. Aber Abweichungen von diesen Prinzipien führen letztlich zur Schwächung der Sicherheit. Insbesondere die Beeinflussung der Darstellung durch die Absenderin kann von einer Angreiferin genutzt werden, um die kognitiven Schwächen der Nutzerinnen auszunutzen.

Die Arbeit bietet eine neue Betrachtungsweise und öffnet neue Wege zur Bekämpfung von Phishing. Dies bedeutet im Umkehrschluss, dass die präsentierten Formalisierungen und Verfahren noch nicht ihren endgültigen Zustand erreicht haben, sondern weiterentwickelt werden können und müssen. Bei vielen Design-Entscheidungen gibt es plausible alternative Möglichkeiten und erst durch eine weitere Nutzung wird die Formalisierung verfestigt. Demzufolge ergeben sich aus der Arbeit eine Vielzahl an weiteren Arbeiten in ganz unterschiedlichen Richtungen.

8.1 Ausblick

Die Arbeit kann auf verschiedene Weisen weitergeführt werden. Dies umfasst die Modellierung, Evaluation von aktuellen Anwendungen, die Konstruktion von neuen Varianten vom vorgestellten Verfahren und anderen Verfahren sowie die Implementierung von dem beschriebenen Verfahren.

8.1.1 Verbesserung vom Formalismus

Die Formalisierung ist ein erster Beitrag und ist damit noch nicht etabliert. Andere Forscherinnen und Nutzerinnen der Formalisierung können diese anpassen, verbessern, korrigieren oder adaptieren. Dies ist ein üblicher und langfristiger Prozess in der Wissenschaft. Modellierungen sind selten bei einem ersten Vorschlag perfekt.

Der Schwerpunkt der Sicherheitsspiele ist die Bestimmung der Herkunft von Nachrichten. Die Sicherheitsspiele lassen sich aber auf diverse andere Probleme adaptieren. Beispielsweise kann das Sicherheitsspiel leicht angepasst auf die Darstellung der Herkunft einer Webseite im Web-Browser übertragen werden. In diesem Fall muss das Gefahrenmodell berücksichtigt werden. Insbesondere die Fälle von Buchstabendrehern bei der Eingabe von URLs sowie das Öffnen eines Links sind zu berücksichtigen. Das Sicherheitsspiel zur Darstellung der Herkunft kann auf die Darstellung von einem Link in einer E-Mail oder Webseite übertragen werden und ist dieser sehr ähnlich.

Der Vergleich von zwei Darstellungen zur Validierung von kryptographischen Schlüsseln durch den Menschen ist ein vielfach betrachtetes Problem und wird in manchen Kommunikationsanwendungen verwendet [7, 131, 135]. Dieses Problem kann mit Sicherheitsspielen modelliert werden. In diesem Fall ist die Aufgabe für den Menschen sehr ähnlich zu der Anfrage `equal`. Folglich muss bei all diesen Anwendungen die menschliche Kognition bei der Konstruktion eines Verfahrens berücksichtigt werden. Azimpourkivi et al. haben einen ersten Versuch zum Vergleich von zwei (Hash)-Darstellungen unternommen [7]. Sie haben ein maschinelles Lernmodell, welches möglichst unterschiedliche Darstellungen für Schlüsselvergleiche konstruiert, entwickelt und dabei einen Datensatz mittels M-Turks erstellt [7]. Dieser Ansatz kann mit dem hier vorgeschlagenen Formalismus und Erkenntnissen aus der Kognitionswissenschaft weiterentwickelt werden.

Die formale Betrachtungsweise von Phishing-Angriffen ermöglicht eine neue Betrachtungsweise auf diese Probleme und damit den Entwurf eines neuen Verfahrens. Denn bei vielen dieser Herausforderungen sind Experimente sehr aufwändig und kostenintensiv. Gleichzeitig sind bei den Experimenten häufig die Varianten der Angriffe sehr eingeschränkt, weil dies mit einem rasanten Anstieg an Studienteilnehmenden einhergehen kann. Beispielsweise benötigten Dechand et al. für ihre Studie zur Erkennung von Angriffen von Hashcodes mehr als 1.000 Studienteilnehmende [37]. Bisher wurde die Ableitung der Herkunft untersucht und in einem nächsten Schritt kann auf ähnliche Weise und als eigenes Spiel die Ausführung einer Aktion formalisiert werden. Dabei können Verfahren zur Darstellung einer URL entwickelt werden oder der Umgang mit bestimmten Aktionen, wie das Zurücksetzen von einem Passwort. Beispielsweise

wäre es möglich, dass bereits bei der Erstellung von einem Account ein Link zum Zurücksetzen hinterlegt wird und dieser nur mit einer späteren Nachricht verwendet wird. In diesem Fall ist es nicht möglich, dass eine Angreiferin einen falschen Link nutzen kann.

Eine weitere Möglichkeit ist die Einführung von Sicherheitsrichtlinien, um bestimmte Aktionen wie das Öffnen von Dateien oder den Klick auf einen Link nur unter bestimmten Bedingungen zu erlauben.

Diese und andere Verfahren erfordern aber das Zusammenwirken von mehreren Systemen und sind deutlich komplexer als das hier vorgeschlagene Verfahren. Aus diesem Grund wurde die Bestimmung der Herkunft weiterverfolgt und vertieft.

8.1.2 Formalisierung der menschen-zentrierten Sicherheit

Die menschen-zentrierte Sicherheit ist insbesondere zwischen IT-Sicherheit sowie Kryptographie und der Psychologie sowie der Kognitionswissenschaft angesiedelt. Bei konstruktiven anwendungsorientierten Vorschlägen zur Verbesserung der Sicherheit müssen beide Perspektiven berücksichtigt werden. Das vorgeschlagene Modell bietet in Bezug auf Phishing-Angriffe eine Möglichkeit hierzu. In der Vergangenheit habe ich bereits eine Modellierung der menschlichen Angreiferin in Bezug auf Mensch-Computer-Authentifikation vorgeschlagen [149]. Im Gegensatz zu dem hier vorgeschlagenen Modell wird bei der Mensch-Computer-Authentifikation der Mensch aus einer Angriffsposition betrachtet. Damit unterscheiden sich diese Modellierungen wesentlich. Perspektivisch ist es wünschenswert, beide Modellierungen zu harmonisieren und zu vereinheitlichen, um daraus eine größere Theorie zu entwickeln. Eine wichtige offene Frage ist dabei, ob es noch weitere fundamental unterschiedliche Modellierungen von Menschen in Bezug auf die IT-Sicherheit gibt und wie diese berücksichtigt werden können.

8.1.3 Varianten vom Verfahren und andere Verfahren

Das vorgestellte Verfahren ist sehr einfach in der Darstellung und der kognitive Hintergrund für die Sicherheit ist aus dem Alltag bekannt sowie wissenschaftlich etabliert. Damit wurde die vorgeschlagene Methode illustriert und verdeutlicht. Es sind aber andere Varianten möglich. Beispielsweise können andere Darstellungsarten wie zum Beispiel künstliche Gesichter oder die Personalisierung der Darstellung zur Verbesserung der Wiedererkennung untersucht werden. Es gibt einen freiverfügbaren Satz[1] von 100.000 künstlichen Gesichtern mit einem einheitlichen Stil. Diese künstlichen Gesichtern können eine Alternative zu richtigen Gesichtern sein. Die Verwendung von künstlichen Gesichter ermöglicht eine einfache Unterscheidung von zufälligen und von Nutzerinnen ausgewählten Gesichtern. Diese Unterscheidung kann hilfreich sein, wenn Nutzerinnen oder Organisationen ein Adressbuch mit Gesichtern pflegen.

Eine weitere Variante ist die Einführung einer minimalen Semantik. Separate Grafiken für die Domain und der lokale Teil einer E-Mail-Adresse bilden eine minimale Semantik. Die grafische Darstellung von

[1] GoogleCartoon: https://google.github.io/cartoonset/

E-Mail-Adressen mit der gleichen Domain kann dann in Teilen gleich dargestellt werden. Offen ist die Art der Darstellung der Domain und ob der gleiche lokale Teil gleich dargestellt werden sollte.

Eine andere Alternative ist die zufällige Zuordnung anzupassen. Eine Möglichkeit ist, die Wahl der Darstellung einer vertrauenswürdigen Dritten-Partei zu überlassen. Diese Überlegungen gibt es in einem RFC-Entwurf von Google und anderen größeren Unternehmen [18].

8.1.4 Bedeutung vom Adressbuch

Das Gefahrenmodell und die Sicherheitsspiele verdeutlichen, dass ein Phishing-Angriff insbesondere durch die Vortäuschung einer legitimen Herkunft charakterisiert wird. Folglich ist es sinnvoll, wenn der Computer die legitimen Herkünfte aus der Vergangenheit kennt und so Nutzerinnen bei der Erkennung unterstützen kann. Eine Herausforderung ist die Beziehung zwischen unterschiedlichen Adressen, wenn diese eigentlich zur selben Person gehören.

Ein gepflegtes Adressbuch kann dabei sehr hilfreich sein. Statt zufällige Gesichter zu einer Herkunft anzuzeigen, kann dann das richtige Gesicht angezeigt werden. Damit wird die Wiedererkennung verbessert. Ein Adressbuch steigert damit nicht nur die Benutzbarkeit, sondern kann die Sicherheit erhöhen. Die Verbesserung der Pflege von einem Adressbuch ist eine unbeachtete Problemstellung. Insbesondere die Verknüpfung von zwei unterschiedlichen Adressen zu einem Adressbucheintrag ist eine sicherheitskritische Entscheidung.

Neben einem klassischen Adressbuch ist die automatisierte Pflege von einem Empfangsbuch aller eingehenden (und ausgehenden) Adressen sinnvoll. Insbesondere bei E-Mails mit einer einseitigen Kommunikation ist es hilfreich zur Erkennung von Angriffen.

Es wurde bisher angenommen, dass vorher bereits eine Kommunikation mit der legitimen Partei erfolgte. Dies muss nicht immer der Fall sein. Beispielsweise konnte die legitimen Partei durch den Besuch der Webseite vertraut sein. In weiteren Arbeiten ist es also denkbar die Vergangenheit von verschiedenen Anwendungen zusammenzuführen. Beispielsweise können die Browser-Historie mit dem E-Mail-Postfach verbunden werden.

Für Organisationen ist es hilfreich, ein gemeinsames Kontaktbuch zu pflegen und bereitzustellen. In diesem Fall kann zu jeder Person innerhalb der Organisation bereits das Gesicht dargestellt werden.

Die vorgestellten Ideen können in weiteren Arbeiten ausgearbeitet und untersucht werden.

8.1.5 Sammlung von Felderfahrungen

In den letzten Jahren gab es unter anderem mit DKIM und SPF Anstrengungen zur Erhöhung der Sicherheit der E-Mail. Bei gängigen E-Mail-Anwendungen wird die Herkunft einer E-Mail auf das $\texttt{FROM}-Feld$ reduziert und andere Informationen nicht berücksichtigt. E-Mail-Anwendungen reduzieren ihre Aufgabe

oftmals auf reine Darstellung und überlassen die Sicherheit den E-Mail-Providern. Die Arbeit zeigt den Bedarf für Veränderungen und das Potential von E-Mail-Anwendungen.

Die Sicherheit des vorgestellten Verfahrens wurde durch wissenschaftliche Erkenntnisse der Kognitionswissenschaft abgeleitet. Dies zeigt, dass das Verfahren und die Prinzipien geeignet sind und weitere Untersuchungen sinnvoll sind. Im nächsten Schritt ist die Einbindung von diesem Verfahren in eine Anwendung nötig. In diesem Schritt ist die Umsetzung von Design und der eigentlichen Anwendung abhängig. In diesem Fall muss das Verfahren erneut untersucht werden und ein Experiment kann sich an dem Spiel 3 orientieren. Durch die Einbettung des Verfahrens in eine Anwendung ist das Ergebnis sehr stark von der Anwendung beeinflusst und lässt sich nicht mehr direkt übertragen. Aus diesem Grund ist diese Ebene der Untersuchung insbesondere durch einen Anbieter einer E-Mail-Anwendung sinnvoll und nötig.

Die Arbeit soll motivieren, die hier vorgeschlagene Darstellung in eine praktische Anwendung zu bringen sowie über Sicherheitsmaßnahmen abseits von Warnungen und Training für Nutzerinnen zu entwickeln. Für IT-Sicherheitsbeauftragte in einer Organisation zeigt die Arbeit, dass ein Adressbuch eine Sicherheitsmaßnahme sein kann und nicht nur die Benutzbarkeit erhöht. Eine Empfehlung an IT-Sicherheitsbeauftragte ist ein gemeinsames Adressbuch mit den Gesichtern von Angehörigen der Organisation zu pflegen und dies gegenüber anderen Interessensgruppen als eine Sicherheitsmaßnahme gegen Phishing und Ransomware zu betrachten.

Glossar

$\mathcal{A}$ ist die Abstraktion der Angreiferin. 39

Gen ist die Initialisierungsmethode um später Nachrichten darzustellen. 39

equal hat zwei Eingaben und $\mathcal{H}$ entscheidet, ob diese gleich sind oder nicht. Diese Anfrage simuliert einen möglichen Wahrnehmungsfehler und orientiert sich an der Aufgabe im Experiment von Mueller und Weidemann [91]. 69

map bezeichnet die Anfrage an $\mathcal{H}$ eine wiedererkannte Darstellung einer Herkunft zuzuordnen. 99

origin bezeichnet die Anfrage an $\mathcal{H}$ nach der Herkunft einer Nachricht. 49

recognize bezeichnet die Anfrage an $\mathcal{H}$, ob eine Darstellung wiederholt vorkommt oder nicht.. 99

meaning bezeichnet die Anfrage an $\mathcal{H}$ zur Interpretation der Darstellung einer Nachricht oder Teile davon. 75

$\mathcal{L}$ bezeichnet den Darstellungsraum zur Darstellung einer Nachricht oder Teile einer Nachricht. 70

Leak umfasst die zusätzlichen Informationen, die einer Angreiferin zur Verfügung gestellt werden. 40

Π setzt sich aus Gen, R und beschreibt ein Verfahren zur Darstellung einer Nachricht oder einem Teil einer Nachricht. 39

legit bezeichnet die Anfrage an $\mathcal{H}$, ob eine Nachricht legitim ist. 40

`sudo` bezeichnet die Anfrage an $\mathcal{H}$ nach der sicherheitskritischen Aktion einer Nachricht. 49

$\mathbf{R}_S$ ist die Darstellung der Herkunft einer Nachricht. 54, 56

$\mathbf{R}$ ist die Darstellung einer Nachricht. 39, 56

Setup ist die Abstraktion zur Erzeugung vom Kontext einer Nachricht. Insbesondere die vergangenen Kommunikation kann dies umfassen.. 43

Σ ist ein Zustand und kann zum Beispiel die vergangene Kommunikation umfassen. 40

$\mathcal{H}$ ist die Abstraktion der menschlichen Entscheidung als *Black-Box*. 42

α ist die Wahrscheinlichkeit, dass Nutzerinnen eine Nachricht akzeptieren. 47

β ist die Erfolgswahrscheinlichkeit einer Angreiferin. 41, 47

ϵ beschreibt die Brauchbarkeit eines Verfahrens und wird durch $\frac{\alpha}{\beta}$ berechnet. 48

γ bezeichnet die sicherheitskritische Aktion in einer Nachricht. 49

φ bezeichnet die Herkunft einer Nachricht. 49

vrf Die Herkunft einer Nachricht wird mit der (kryptographischen) Funktion vrf eindeutig und korrekt bestimmt. Im Kontext der E-Mail kann dies beispielsweise eine digitale Signatur mittels DKIM, S/MIME oder PGP sein. 98

DKIM ist die Abkürzung für *Domain Key Identified Mail* und ist im RFC6376 [75] spezifiziert. Mit diesem Protokoll kann ein Server E-Mails signieren. Die Herkunft einer E-Mail kann mit diesem Protokoll überprüft werden. Gemeinsam mit SPF bildet es die Grundlage für DMARC. 20, 133, 134

DMARC ist die Abkürzung für *Domain-based Message Authentication, Reporting, and Conformance* und ist im RFC7489 [76] spezifiziert. Es vereint die DKIM und SPF. Es bietet eine Entscheidungshilfe zum Umgang mit E-Mails, welche nach einem der beiden Protokoll eine zweifelhafte Herkunft haben. 21, 133, 134

DNS ist die Abkürzung für *Domain Name System* und ist unter anderem im RFC8446 [109]. Es ermöglicht die Auflösung einer *Domain* zu einer IP-Adresse und damit zu einem Server. Gleichzeitig ist es möglich, in einem DNS-Eintrag weitere Informationen zu hinterlegen. 15, 134

FPR ist die Abkürzung für *false positive rate*. In den verwendeten Spielen bildet dies die falsche Akzeptanzrate oder die falsche Wiedererkennungsrate. 47

MSE ist die Abkürzung für den mittlere quadratische Fehler. 105

OpenPGP ist die Abkürzung für *Open Pretty Good Privacy* und ist im RFC3156[114] und RFC4480[78]. Es ermöglicht die Ende-zu-Ende-Verschlüsselung einer E-Mail sowie die digitale Signierung einer E-Mail. 22

PKI ist die Abkürzung für eine (kommerziellen) Infrastruktur für öffentlicher Schlüssel und wird unter anderem für DKIM oder S/MIME genutzt. 21

RanVisCol bezeichnet ein Kognitionsspiel, in dem zufällige Darstellungen vom Menschen als gleich wahrgenommen werden. 70

S/MIME ist die Abkürzung für *Secure/Multipurpose Internet Mail Extensions* und ist im RFC3211 [54]. Es ermöglicht die Ende-zu-Ende-Verschlüsselung einer E-Mail sowie die digitale Signierung einer E-Mail. 21, 133

SelVisCol bezeichnet ein Kognitionsspiel, in dem eine Angreiferin zu einer gegeben Darstellung eine andere Darstellung finden muss und beide werden vom Menschen als gleich wahrgenommen. 70

SPF ist die Abkürzung das *Sender Policy Framework* und ist im RFC7208 [69] spezifiziert. In diesem Protokoll kann ein E-Mail-Server die IP-Adressen, welche zum Versand von E-Mails für eine bestimmte Domain genutzt werden, über einen DNS-Eintrag veröffentlichen. Die Herkunft einer E-Mail kann mit diesem Protokoll in Teilen überprüft werden. Gemeinsam mit DKIM bildet es die Grundlage für DMARC. 19, 133

TLS ist die Abkürzung für *Transport Layer Security* und ist im RFC6376 [75] spezifiziert. Mit diesem Protokoll kann zwischen zwei Servern oder zwischen einem Server und einem Client ein verschlüsselter Kanal gestartet werden. TLS hat viele Einsatzmöglichkeiten und wird unter anderem bei der E-Mail oder bei Webseiten verwendet . 18

TPR ist die Abkürzung für *true positive rate*. 47, 103

VisCol bezeichnet ein Kognitionsspiel, in dem eine Angreiferin zwei unterschiedliche Darstellungen finden muss, welche vom Menschen als gleich wahrgenommen werden. 70

Literaturverzeichnis

[1] Andre Adelsbach, Sebastian Gajek, and Jörg Schwenk. Visual spoofing of SSL protected web sites and effective countermeasures. In *Information Security Practice and Experience: First International Conference, Singapore*. Springer, April 2005.

[2] Pieter Agten, Wouter Joosen, Frank Piessens, and Nick Nikiforakis. Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In *NDSS'15*. Internet Society, 2015.

[3] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *USENIX Security'13*. USENIX, 2013.

[4] Sara Albakry, Kami Vaniea, and Maria K. Wolters. What is This URL's Destination? Empirical Evaluation of Users' URL Reading.

[5] Tarfah Alrashed, Chia-Jung Lee, Peter Bailey, Christopher Lin, Milad Shokouhi, and Susan Dumais. Evaluating user actions as a proxy for email significance. In *WWW'19*. ACM, 2019.

[6] Kholoud Althobaiti, Nicole Meng, and Kami Vaniea. I Don't Need an Expert! Making URL Phishing Features Human Comprehensible. In *CHI'21*. ACM, 2021.

[7] Mozhgan Azimpourkivi, Umut Topkara, and Bogdan Carbunar. Human distinguishable visual key fingerprints. In *USENIX Security'20*, 2020.

[8] Wilma Bainbridge, Phillip Isola, Idan Blank, and Aude Oliva. Establishing a database for studying human face photograph memory. In *Proceedings of the annual meeting of the cognitive science society*, volume 34, 2012.

[9] Wilma A Bainbridge. Memorability: How what we see influences what we remember. In *Psychology of learning and motivation*, volume 70. Elsevier, 2019.

[10] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142(4), 2013.

[11] David Basin, Saa Radomirovic, and Lara Schmid. Modeling human errors in security protocols. In *IEEE 29th Computer Security Foundations Symposium*. IEEE, 2016.

[12] Bernhard Beckert and Gerd Beuster. A method for formalizing, analyzing, and verifying secure user interfaces. In *Formal Methods and Software Engineering: 8th International Conference on Formal Engineering Methods, Macao, China*. Springer, November 2006.

[13] Zinaida Benenson, Gabriele Lenzini, Daniela Oliveira, Simon Parkin, and Sven Uebelacker. Maybe Poor Johnny Really Cannot Encrypt: The Case for a Complexity Theory for Usable Security. ACM, 2015.

[14] Frank Bentley, Nediyana Daskalova, and Nazanin Andalibi. "If a Person is Emailing You, It Just Doesn'T Make Sense": Exploring Changing Consumer Behaviors in Email. In *CHI'17*.

[15] Abhay Bhushan, Ken Pogran, Ray Tomlinson, and Jim White. Standardizing Network Mail Headers. RFC 561, September 1973.

[16] Robert Biddle, Sonia Chiasson, and P.C. Van Oorschot. Graphical Passwords: Learning from the First Twelve Years. 2012.

[17] Hugo LJ Bijmans, Tim M Booij, Anneke Schwedersky, Aria Nedgabat, and Rolf van Wegberg. Catching Phishers By Their Bait: Investigating the Dutch Phishing Landscape through Phishing Kit Detection. In *USENIX Security'21*. USENIX, 2021.

[18] Seth Blank, Peter Goldstein, Thede Loder, Terry Zink, Marc Bradshaw, and Alex Brotman. Brand Indicators for Message Identification (BIMI). Internet-Draft draft-brand-indicators-for-message-identification-03, Internet Engineering Task Force, April 2023. Work in Progress.

[19] David B Boles and John E Clifford. An upper-and lowercase alphabetic similarity matrix, with derived generation similarity values. *Behavior Research Methods, Instruments, & Computers*, 21(6), 1989.

[20] Joseph Bonneau and Stuart Schechter. Towards reliable storage of 56-bit secrets in human memory. In *USENIX Security'14*. USENIX, 2014.

[21] Glencora Borradaile, Kelsy Kretschmer, Michele Gretes, and Alexandria LeClerc. The Motivated Can Encrypt (Even with PGP). *PETS'21*, 2021.

[22] Jürgen Bortz and Gustav A Lienert. *Kurzgefasste Statistik für die klinische Forschung: Leitfaden für die verteilungsfreie Analyse kleiner Stichproben*. Springer-Verlag, 2008.

[23] Sacha Brostoff and M. Angela Sasse. Are Passfaces More Usable Than Passwords? A Field Trial Investigation. In *People and Computers XIV — Usability or Else!*, London, 2000. Springer.

[24] Paolo Buono, Giuseppe Desolda, Francesco Greco, and Antonio Piccinno. Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings. In *Extended Abstracts of CHI'23*. ACM, 2023.

[25] Zoya Bylinskii, Lore Goetschalckx, Anelise Newman, and Aude Oliva. Memorability: An image-computable measure of information utility. In *Human Perception of Visual Information*. Springer, 2022.

[26] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. NoPhish app evaluation: lab and retention study. In *Workshop on Usable Security'15*. Internet Society, 2015.

[27] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. IEEE, 2013.

[28] Stuart K Card, Thomas P Moran, and Allen Newell. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7), 1980.

[29] Angus F Chapman, Hannah Hawkins-Elder, and Tirta Susilo. How robust is familiar face recognition? A repeat detection study of more than 1000 faces. *Royal Society open science*, 5, 2018.

[30] Jianjun Chen, Vern Paxson, and Jian Jiang. Composition Kills: A Case Study of Email Sender Authentication. In *USENIX Security 20*. USENIX, August 2020.

[31] Robert B Cialdini and Robert B Cialdini. *Influence: The psychology of persuasion*, volume 55. Collins New York, 2007.

[32] Charles Clifton Jr, Fernanda Ferreira, John M Henderson, Albrecht W Inhoff, Simon P Liversedge, Erik D Reichle, and Elizabeth R Schotter. Eye movements in reading and information processing: Keith Rayner's 40 year legacy. *Journal of Memory and Language*, 86, 2016.

[33] DWJ Corcoran. An acoustic factor in letter cancellation. *Nature*, 210(5036), 1966.

[34] Lorrie F Cranor. A framework for reasoning about the human in the loop. Advanced Computing Systems Professional and Technical Association, 2008.

[35] Tom Cuchta, Brian Blackwood, Thomas R Devine, Robert J Niichel, Kristina M Daniels, Caleb H Lutjens, Sydney Maibach, and Ryan J Stephenson. Human risk factors in cybersecurity. In *Proceedings of the 20th annual SIG conference on information technology education*, pages 87–92, 2019.

[36] Darren Davis, Fabian Monrose, and Michael K. Reiter. On User Choice in Graphical Password Schemes. USENIX, 2004.

[37] Sergej Dechand, Dominik Schürmann, Karoline Busse, Yasemin Acar, Sascha Fahl, and Matthew Smith. An Empirical Study of Textual Key-Fingerprint Representations. In *USENIX Security'16*. USENIX, 2016.

[38] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why Phishing Works. ACM, 2006.

[39] Rachna Dhamija and J Doug Tygar. The battle against phishing: Dynamic security skins. In *SOUPS'05*, 2005.

[40] Dotan Di Castro, Zohar Karnin, Liane Lewin-Eytan, and Yoelle Maarek. You've got mail, and here is what you could do with it! Analyzing and predicting actions on email messages. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining'16*, 2016.

[41] Bruno Dujardin, Jef Van den Ende, Alfons Van Gompel, Jean-Pierre Unger, and Patrick Van der Stuyft. Likelihood ratios: a real improvement for clinical decision making? *European journal of epidemiology*, 10, 1994.

[42] Zakir Durumeric, David Adrian, Ariana Mirian, James Kasten, Elie Bursztein, Nicolas Lidzborski, Kurt Thomas, Vijay Eranti, Michael Bailey, and J. Alex Halderman. Neither Snow Nor Rain Nor MITM...: An Empirical Analysis of Email Delivery Security. In *IMC'15*.

[43] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *CHI'08*. ACM, 2008.

[44] Carl Ellison. Ceremony design and analysis. *Cryptology EPrint Archive*, 2007.

[45] Danyel Fisher, AJ Brush, Eric Gleave, and Marc A Smith. Revisiting Whittaker & Sidner's email overload ten years later. In *CSCW'06*. ACM, 2006.

[46] Ian D. Foster, Jon Larson, Max Masich, Alex C. Snoeren, Stefan Savage, and Kirill Levchenko. Security by Any Other Name: On the Effectiveness of Provider Based Email Security. In *CCS'15*.

[47] Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig, Christian Reuter, Alexander Benlian, and Joachim Vogt. SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research. In *SOUPS'21*. USENIX, August 2021.

[48] Sebastian Gajek, Mark Manulis, Ahmad-Reza Sadeghi, and Jörg Schwenk. Provably secure browser-based user-aware mutual authentication over TLS. In *CCS'08*. ACM, 2008.

[49] Isabel Gauthier, Pawel Skudlarski, John C Gore, and Adam W Anderson. Expertise for cars and birds recruits brain areas involved in face recognition. *Nature neuroscience*, 3(2), 2000.

[50] Lore Goetschalckx, Pieter Moors, and Johan Wagemans. Image memorability across longer time intervals. *Memory*, 26(5), 2018.

[51] Shafi Goldwasser and Yael Tauman Kalai. Cryptographic assumptions: A position paper. In *Theory of Cryptography: 13th International Conference, Tel Aviv, Israel*. Springer, 2016.

[52] Kristen K Greene, Michelle Steves, Mary F Theofanos, and Jennifer Kostick. User context: an explanatory variable in phishing susceptibility. In *Workshop Usable Security'18*. Internet Society, 2018.

[53] Catherine Grevet, David Choi, Debra Kumar, and Eric Gilbert. Overload is overloaded: email in the age of Gmail. In *CHI'14*. ACM, 2014.

[54] Peter Gutmann. Password-based Encryption for CMS. RFC 3211, December 2001.

[55] Harry Halpin. SoK: why Johnny can't fix PGP standardization. In *ARES'20*, 2020.

[56] Alice F Healy. Detection errors on the word the: evidence for reading units larger than letters. *Journal of Experimental Psychology: Human Perception and Performance*, 2, 1976.

[57] Alice F Healy. The effects of visual similarity on proofreading for misspellings. *Memory & Cognition*, 9(5), 1981.

[58] Leslie Henderson. Writing systems and reading processes. In *Orthographies and reading*. Routledge, 2017.

[59] Amir Herzberg and Ahmad Jbara. Security and Identification Indicators for Browsers against Spoofing and Phishing Attacks. 2008.

[60] Paul E. Hoffman. SMTP Service Extension for Secure SMTP over Transport Layer Security. RFC 3207, February 2002.

[61] Nicholas J. Hopper and Manuel Blum. Secure human identification protocols. In *ASIACRYPT*, 2001.

[62] Hang Hu and Gang Wang. End-to-End Measurements of Email Spoofing Attacks. In *USENIX Security'18*. USENIX, 2018.

[63] Danny Yuxing Huang, Maxwell Matthaios Aliapoulios, Vector Guo Li, Luca Invernizzi, Elie Bursztein, Kylie McRoberts, Jonathan Levin, Kirill Levchenko, Alex C Snoeren, and Damon McCoy. Tracking ransomware end-to-end. In *IEEE Symposium on Security and Privacy'18*. IEEE, 2018.

[64] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *CVPR 2011*. IEEE, 2011.

[65] Xiaofeng Zheng Minglei Guo Chaoyi Lu Baojun Liu Yuxuan Zhao Shuang Hao Haixin Duan Qingfeng Pan Kaiwen Shen, Chuhan Wang and Min Yang. Weak Links in Authentication Chains: A Large-scale Analysis of Email Sender Spoofing Attacks. In *USENIX Security'21*, 2021.

[66] Jonathan Katz and Yehuda Lindell. *Introduction to modern cryptography*. CRC press, 2020.

[67] Mannat Kaur, Michel van Eeten, Marijn Janssen, Kevin Borgolte, and Tobias Fiebig. Human factors in security research: Lessons learned from 2008-2018. *arXiv preprint arXiv:2103.13287*, 2021.

[68] Aditya Khosla, Akhil S Raju, Antonio Torralba, and Aude Oliva. Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[69] Scott Kitterman. Sender Policy Framework (SPF) for Authorizing Use of Domains in Email, Version 1. RFC 7208, April 2014.

[70] Dr. John C. Klensin. Simple Mail Transfer Protocol. RFC 5321, October 2008.

[71] Neal Koblitz and Alfred Menezes. Critical perspectives on provable security: Fifteen years of another look papers, 2019.

[72] Paul C Kocher. Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In *CRYPTO'96, Santa Barbara, California, USA*. Springer, 1996.

[73] Viktor Krammer. Phishing defense against IDN address spoofing attacks. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services*, 2006.

[74] Kat Krol, Jonathan M. Spring, Simon Parkin, and M. Angela Sasse. Towards Robust Experimental Design for User Studies in Security and Privacy. In *The LASER Workshop: Learning from Authoritative Security Experiment Results*, San Jose, CA, May 2016. USENIX.

[75] Murray Kucherawy, Dave Crocker, and Tony Hansen. DomainKeys Identified Mail (DKIM) Signatures. RFC 6376, September 2011.

[76] Murray Kucherawy and Elizabeth Zwicky. Domain-based Message Authentication, Reporting, and Conformance (DMARC). RFC 7489, March 2015.

[77] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny Not to Fall for Phish. 2010.

[78] Paul Kyzivat, Vijay K. Gurbani, Henning Schulzrinne, and Jonathan Rosenberg. RPID: Rich Presence Extensions to the Presence Information Data Format (PIDF). RFC 4480, July 2006.

[79] Daniele Lain, Kari Kostiainen, and Srdjan Čapkun. Phishing in organizations: Findings from a large-scale and long-term study. In *IEEE Symposium on Security and Privacy'22*. IEEE, 2022.

[80] Joscha Lausch, Oliver Wiese, and Volker Roth. What is a secure email? In *European Workshop on Usable Security'17*, 2017.

[81] Ada Lerner, Eric Zeng, and Franziska Roesner. Confidante: Usable encrypted email: A case study with lawyers and journalists. In *IEEE European Symposium on Security and Privacy'17*. IEEE, 2017.

[82] Vladimir I Levenshtein et al. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[83] Stephen J Lupker. Visual word recognition: Theories and findings. *The science of reading: A handbook*, 2005.

[84] Peter Mayer, Damian Poddebniak, Konstantin Fischer, Marcus Brinkmann, Juraj Somorovsky, Angela Sasse, Sebastian Schinzel, and Melanie Volkamer. I don't know why I check this...'' - Investigating Expert Users' Strategies to Detect Email Signature Spoofing Attacks. In *SOUPS'22*, 2022.

[85] Susan E McGregor, Elizabeth Anne Watkins, Mahdi Nasrullah Al-Ameen, Kelly Caine, and Franziska Roesner. When the Weakest Link is Strong: Secure Collaboration in the Case of the Panama Papers. In *USENIX Security'17*. USENIX, 2017.

[86] Alexey Melnikov, Tony Hansen, and Chris Newman. Internationalized Delivery Status and Disposition Notifications. RFC 6533, February 2012.

[87] Alexey Melnikov and Barry Leiba. Internet Message Access Protocol (IMAP) - Version 4rev2. RFC 9051, August 2021.

[88] Laura Mickes, Heather D Flowe, and John T Wixted. Receiver operating characteristic analysis of eyewitness memory: comparing the diagnostic accuracy of simultaneous versus sequential lineups. *Journal of Experimental Psychology: Applied*, 18(4), 2012.

[89] Tyler Moore and Benjamin Edelman. Measuring the perpetrators and funders of typosquatting. In *Financial Cryptography and Data Security: 14th International Conference, Tenerife, Canary Islands*. Springer, January 2010.

[90] Charles A Morgan III, Gary Hazlett, Madelon Baranoski, Anthony Doran, Steven Southwick, and Elizabeth Loftus. Accuracy of eyewitness identification is significantly associated with performance on a standardized test of face recognition. *International Journal of Law and Psychiatry*, 30(3), 2007.

[91] Shane T Mueller and Christoph T Weidemann. Alphabetic letter identification: Effects of perceivability, similarity, and bias. *Acta psychologica*, 139(1), 2012.

[92] Jens Müller, Marcus Brinkmann, Damian Poddebniak, Hanno Böck, Sebastian Schinzel, Juraj Somorovsky, and Jörg Schwenk. "Johnny, you are fired!" – Spoofing OpenPGP and S/MIME Signatures in Emails. In *USENIX Security'19*, 2019.

[93] Jens Müller, Marcus Brinkmann, Damian Poddebniak, Sebastian Schinzel, and Jörg Schwenk. Re: What's Up Johnny? Covert Content Attacks on Email End-to-End Encryption. In *Applied Cryptography and Network Security: 17th International Conference, Bogota, Colombia*. Springer, 2019.

[94] Jens Müller, Marcus Brinkmann, Damian Poddebniak, Sebastian Schinzel, and Jörg Schwenk. Mailto: Me your secrets. on bugs and features in email end-to-end encryption. In *2020 IEEE Conference on Communications and Network Security*. IEEE, 2020.

[95] Douglas L Nelson, Valerie S Reed, and John R Walling. Pictorial superiority effect. *Journal of experimental psychology: Human learning and memory*, 2(5), 1976.

[96] Chris Newman. Using TLS with IMAP, POP3 and ACAP. RFC 2595, June 1999.

[97] Adam Oest, Yeganeh Safaei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Kevin Tyers. Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists. In *IEEE Symposium on Security and Privacy'19*. IEEE, 2019.

[98] National Institute of Standards and Technology. An Introduction to Information Security. Technical Report NIST Special Publication 800-12 Revision 1, U.S. Department of Commerce, Washington, D.C., June 2017.

[99] Justin Petelka, Yixin Zou, and Florian Schaub. Put Your Warning Where Your Link Is: Improving and Evaluating Email Phishing Warnings. In *CHI'19*. ACM, 2019.

[100] Damian Poddebniak, Christian Dresen, Jens Müller, Fabian Ising, Sebastian Schinzel, Simon Friedberger, Juraj Somorovsky, and Jörg Schwenk. Efail: Breaking S/MIME and OpenPGP Email Encryption using Exfiltration Channels. In *USENIX Security'18*, 2018.

[101] Jonathan B. Postel. Simple Mail Transfer Protocol. RFC 788, November 1981.

[102] Florian Quinkert, Tobias Lauinger, William Robertson, Engin Kirda, and Thorsten Holz. It's not what it looks like: Measuring attacks and defensive registrations of homograph domains. In *IEEE Conference on Communications and Network Security*. IEEE, 2019.

[103] Kathleen Rastle. Visual word recognition. In *Neurobiology of language*. Elsevier, 2016.

[104] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3), 1998.

[105] Keith Rayner. The gaze-contingent moving window in reading: Development and review. *Visual Cognition*, 22(3-4), 2014.

[106] Keith Rayner, Elizabeth R Schotter, Michael EJ Masson, Mary C Potter, and Rebecca Treiman. So much to read, so little time: How do we read, and can speed reading help? *Psychological Science in the Public Interest*, 17(1), 2016.

[107] Keith Rayner, Sarah J White, Rebecca L Johnson, and Simon P Liversedge. Raeding wrods with jubmled lettres: there is a cost. *Psychological science*, 2006.

[108] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In *CHI'18*. ACM, 2018.

[109] E. Rescorla. RFC 8446: The Transport Layer Security (TLS) Protocol Version 1.3, 2018.

[110] Pete Resnick. Internet Message Format. RFC 2822, April 2001.

[111] Pete Resnick. Internet Message Format. RFC 5322, October 2008.

[112] Joshua Reynolds, Adam Bates, and Michael Bailey. Equivocal URLs: Understanding the Fragmented Space of URL Parser Implementations. In *ESORICS'22, Copenhagen, Denmark*, September.

[113] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. Measuring Identity Confusion with Uniform Resource Locators. In *CHI'20*. ACM, 2020.

[114] Thomas Roessler, Michael Elkins, Raph Levien, and Dave Del Torto. MIME Security with OpenPGP. RFC 3156, August 2001.

[115] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. We're on the Same Page A Usability Study of Secure Email Using Pairs of Novice Users. In *CHI'16*. ACM, 2016.

[116] Scott Ruoti, Jeff Andersen, Travis Hendershot, Daniel Zappala, and Kent Seamons. Private Webmail 2.0: Simple and Easy-to-Use Secure Email. In *UIST'16*. ACM, 2016.

[117] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy van der Horst, and Kent Seamons. Confused Johnny: When Automatic Encryption Leads to Confusion and Mistakes. In *SOUPS'13*. ACM, 2013.

[118] Jörg Schwenk, Marcus Brinkmann, Damian Poddebniak, Jens Müller, Juraj Somorovsky, and Sebastian Schinzel. Mitigation of Attacks on Email End-to-End Encryption. ACM, 2020.

[119] Kaiwen Shen, Chuhan Wang, Minglei Guo, Xiaofeng Zheng, Chaoyi Lu, Baojun Liu, Yuxuan Zhao, Shuang Hao, Haixin Duan, Qingfeng Pan, and Min Yang. Weak Links in Authentication Chains: A Large-scale Analysis of Email Sender Spoofing Attacks. In *USENIX Security'21)*. USENIX, August 2021.

[120] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *CHI'10*. ACM, 2010.

[121] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *SOUPS'07*, 2007.

[122] Robert W. Shirey. Internet Security Glossary, Version 2. RFC 4949, August 2007.

[123] Camelia Simoiu, Ali Zand, Kurt Thomas, and Elie Bursztein. Who is targeted by email-based phishing and malware? measuring factors that differentiate risk. In *IMC'20*, 2020.

[124] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[125] Geoffrey Simpson, Tyler Moore, and Richard Clayton. Ten years of attacks on companies using visual impersonation of domain names. In *APWG Symposium on Electronic Crime Research'20*. IEEE, 2020.

[126] Lionel Standing, Jerry Conezio, and Ralph Norman Haber. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. *Psychonomic science*, 19(2), 1970.

[127] Robert J Sternberg, Karin Sternberg, and Jeff Mio. *Cognitive psychology*. Wadsworth, 2012.

[128] Christian Stransky, Oliver Wiese, Volker Roth, Yasemin Acar, and Sascha Fahl. 27 Years and 81 Million Opportunities Later: Investigating the Use of Email Encryption for an Entire University. In *IEEE Symposium on Security and Privacy'22*. IEEE, May 2022.

[129] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *USENIX Security'09*. USENIX, 2009.

[130] Janos Szurdi, Balazs Kocso, Gabor Cseh, Jonathan Spring, Mark Felegyhazi, and Chris Kanich. The long "taile" of typosquatting domain names. In *USENIX Security'14*. USENIX, 2014.

[131] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can unicorns help users compare crypto key fingerprints? In *CHI'17*. ACM, 2017.

[132] Ke Tian, Steve T. K. Jan, Hang Hu, Danfeng Yao, and Gang Wang. Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild. In *IMC'18*. ACM, 2018.

[133] Tim Valentine. An evaluation of the Passface personal authentication system. Technical report, Technical Report, Goldsmiths College, University of London, 1998.

[134] Amber van der Heijden and Luca Allodi. Cognitive Triaging of Phishing Attacks. In *USENIX Security'19)*, Santa Clara, CA, August 2019. USENIX.

[135] Elham Vaziripour, Justin Wu, Mark O'Neill, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala. Is that you, Alice? A usability study of the authentication ceremony of secure messaging applications. In *SOUPS'17*, 2017.

[136] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of torpedo: Tooltip-powered phishing email detection. *Computers & Security*, 71, 2017.

[137] Artemij Voskobojnikov, Oliver Wiese, Masoud Mehrabi Koushki, Volker Roth, and Konstantin (Kosta) Beznosov. The U in Crypto Stands for Usable: An Empirical Study of User Experience with Mobile Cryptocurrency Wallets. ACM, 2021.

[138] Chuhan Wang, Kaiwen Shen, Minglei Guo, Yuxuan Zhao, Mingming Zhang, Jianjun Chen, Baojun Liu, Xiaofeng Zheng, Haixin Duan, Yanzhong Lin, and Qingfeng Pan. A Large-scale and Longitudinal Measurement Study of DKIM Deployment. In *USENIX Security'22*, 2022.

[139] Yi-Min Wang, Doug Beck, Jeffrey Wang, Chad Verbowski, and Brad Daniels. Strider Typo-Patrol: Discovery and Analysis of Systematic Typo-Squatting. *SRUTI*, 6, 2006.

[140] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L Mazurek, Manya Sleeper, and Kurt Thomas. SoK: A framework for unifying at-risk user research. In *IEEE Symposium on Security and Privacy'22*. IEEE, 2022.

[141] C.A. Weaver and A.E. Holmes. Psychology of reading. In V.S. Ramachandran, editor, *Encyclopedia of Human Behavior*. Academic Press, San Diego, second edition edition, 2012.

[142] Joel Weinberger and Adrienne Porter Felt. A week to remember: The impact of browser warning storage policies. In *SOUPS'16*, 2016.

[143] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In *CHI'19*. ACM, 2019.

[144] Dirk Wentura and Christian Frings. *Kognitive Psychologie*. Springer, 2012.

[145] Steve Whittaker and Candace Sidner. Email overload: exploring personal information management of email. In *CHI'1996*. ACM, 1996.

[146] Thomas D Wickens. *Elementary signal detection theory*. Oxford university press, 2001.

[147] Oliver Wiese, Joscha Lausch, Jakob Bode, and Volker Roth. Beware the Downgrading of Secure Electronic Mail. *Workshop on Socio-Technical Aspects in Security and Trust*, 2018.

[148] Oliver Wiese and Volker Roth. Pitfalls of Shoulder Surfing Studies. In *Workshop on Usable Security'15*. Internet Society, 2015.

[149] Oliver Wiese and Volker Roth. See You next Time: A Model for Modern Shoulder Surfers. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 2016.

[150] Mohammad Younesi and Yalda Mohsenzadeh. Facememnet: Predicting face memorability with deep neural networks. 2022.

[151] Sijie Zhuo, Robert Biddle, Yun Sing Koh, Danielle Lottridge, and Giovanni Russello. SoK: Human-centered phishing susceptibility. *ACM Transactions on Privacy and Security*, 2022.

Anhang A

Über den Autor

Oliver Wiese studierte von 2009 bis 2016 an der Freien Universität Berlin Informatik mit Mathematik im Nebenfach. Der Schwerpunkt des Studiums waren Sicherheit sowie theoretische Informatik. Während seines Studiums wurde er durch das Deutschlandstipendium gefördert. Seine Masterarbeit wurde mit dem dritten Platz beim CAST-Förderpreis für IT-Sicherheit ausgezeichnet und die dazugehörige Publikation wurde auf der MobileHCI'16 als eine von zwei Publikationen mit einem *Best Paper Award* ausgezeichnet [149].

Von 2011 bis 2015 hat er als Tutor am Institut für Informatik eine Vielzahl an verschiedenen Lehrveranstaltungen betreut. Als studentische Hilfskraft in der Arbeitsgruppe von Prof. Dr-Ing. Volker Roth forschte er im Bereich der Sicherheit von Authentifizierungsverfahren.

Von 2016 bis 2023 war er wissenschaftlicher Mitarbeiter in der Arbeitsgruppe von Prof. Dr-Ing. Volker Roth. Im vom BMBF geförderten Projekt Enzevalos forschte er von 2016 bis 2018 gemeinsam mit zwei studentischen Hilfskräften zur Benutzbarkeit von Ende-Zu-Ende-Verschlüsselung und entwickelte mit der Letterbox eine E-Mail-App mit Ende-Zu-Ende-Verschlüsselung auf iOS. Er führte in der Zeit verschiedene Lehrveranstaltungen selbstständig durch und betreute sowohl Bachelor- als auch Masterstudierende bei ihrer Abschlussarbeit.

Zwischen Mai und August 2019 untersuchte er bei einem Forschungsaufenthalt bei Prof. Kosta Beznosov an der University of British Columbia die Benutzbarkeit von *Cryptowallets*. Die Forschungsergebnisse wurden auf der CHI'21 veröffentlicht und mit einer honorable Mention ausgezeichnet [137].

Er war bei der CHI'17 und CHI'18 externer Reviewer und im PC der EuroUSec'21 und vertrat die Statusgruppe der wissenschaftlichen Mitarbeitenden im Fakultäts- und Institutsrat sowie im Prüfungsausschuss. Neben seiner akademischen Arbeit hat er den Verein Collegium Academicum Berlin gegründet und fördert zu dem gemeinsam mit ehemaligen Kommilitonen über diesen Verein ein Deutschlandstipendium. Er ist als Jugendleiter im Alpenverein Sektion Berlin aktiv und seit 2022 Vorstandsvorsitzender vom Ruderverein Märkischer Wassersport.

Seine bisherigen sieben Publikationen wurden in der Vergangenheit mehr als 100 mal zitiert[1].

[1]Eine Übersicht der Publikationen und Referenzen findet sich hier: https://scholar.google.com/citations?user=bMJldNwAAAAJ&hl=de&authuser=1

Anhang B

Zusammenfassung der Ergebnisse

Die Beiträge dieser Arbeit sind zusammengefasst wie folgt:

1. Es wird ein formales Modell entwickelt, welches die Sicherheit gegen Phishing-Angriffe einer Darstellung einer E-Mail in Beziehung zu kognitiven Eigenschaften von Menschen stellt. In diesem Modell werden die Darstellung und die menschliche Entscheidung abstrakt betrachtet. Dies ermöglicht eine Herleitung und Begründung der Sicherheit gegen Phishing-Angriffe.
2. Es werden vier verschiedene Sicherheitsspiele formalisiert und die Beziehung zwischen den Sicherheitsspielen aufgezeigt.
3. Es werden allgemeine Heuristiken zur frühzeitigen Erkennung von potentiell unsicheren Darstellungen aufbauend auf kognitiven Einschränkungen von Menschen entwickelt.
4. Die Heuristiken werden im Kontext von E-Mail-Adressen angewendet und Gefahren bei der Wahrnehmung durch Nutzerinnen werden durch Erkenntnisse aus der Kognitionswissenschaft gezeigt. In einer Datenanalyse von tatsächlichen Phishing-E-Mails werden diese Heuristiken auf E-Mail-Adressen angewendet und die Beziehung zu Einschränkungen von Menschen beim Lesen von Zeichenketten zeigt.
5. Es wird eine Darstellung einer Herkunft einer E-Mail mittels zufällig gewählter Gesichter entwickelt.
6. Die Sicherheit dieser Darstellung wird durch die Ergebnisse aus Experimenten der Kognitionswissenschaft begründet.
7. Die Modellierung und das vorgeschlagene Verfahren werden mit den Ergebnissen von verschiedenen wissenschaftlichen Forschungszweigen (Kryptographie, IT-Sicherheit, Menschen-zentrierte Sicherheit) zu Phishing verglichen. Es werden weitere neue Forschungsrichtungen aufgezeigt und die Bedeutung von einem Adressbuch für die Sicherheit hervorgehoben.

Anhang C

Reproduktion der Forschungsergebnisse

Die Python-Skripte zur Analyse der Zeichenketten und der Gesichter sind unter http://dx.doi.org/10.17169/refubium-41698 verfügbar. Die genutzten Daten sind nicht frei verfügbar, aber sind für wissenschaftliche Arbeiten zugänglich. Der Zugriff auf den Datensatz von Phishing-Angriffen kann beim Cambridge CyberCrime Center (CCCC) unter https://www.cambridgecybercrime.uk/datasets.html beantragt werden. Der Zugriff auf den Datensatz mit den 10k-Gesichtern kann beim Wilma Bainbridge Lab unter https://www.wilmabainbridge.com/facememorability2.html beantragt werden. Das gelernte Modell zur Klassifikation ist nicht öffentlich zugänglich, weil die Gefahr bestehen kann, dass die geschützten Daten extrahiert werden könnten.