Usability von CAPTCHA-Systemen*

Stefan Penninger, Stefan Meier
Universität Regensburg
{stefan.penninger,stefan1.meier}@ur.de

Hannes Federrath
Universität Hamburg
federrath@informatik.uni-hamburg.de

Abstract: CAPTCHA-Systeme sind weit verbreitete Schutzsysteme, um auf Online-Plattformen menschliche Benutzer von automatischen Bots zu unterscheiden. Dabei kommen verschiedene Varianten zum Einsatz, die sich in Art und Interaktionsmodus sowie im Schwierigkeitsgrad der Lösung unterscheiden. In der vorliegenden Studie werden Kriterien der Benutzbarkeit von CAPTCHA-Systemen aufgestellt. Zudem werden fünf typische CAPTCHA-Implementierungen im Bezug auf ihre Gebrauchstauglichkeit mit Hilfe einer Benutzerstudie empirisch untersucht. Während sich Math-CAPTCHAs unter den getesteten Alternativen in den zugrunde gelegten Kriterien der Benutzbarkeit überlegen zeigen, muss unter Einbeziehung von Sicherheitskriterien das klassische Bild-CAPTCHA nach wie vor als das zuverlässigste Mittel der Mensch-Maschine-Unterscheidung gelten.

1 Motivation

CAPTCHAs sind automatisierte Turing-Tests (*Completely Automated Public Turing Test to tell Computers and Humans Apart*), welche auf Webseiten verwendet werden, um menschliche Nutzer von automatischen Skripten zu unterscheiden. Dabei handelt es sich um eine Form von Challenge-Response-Tests: Es werden Aufgaben gestellt, die für den menschlichen Nutzer möglichst einfach zu beantworten sein sollen, sich jedoch nicht effizient durch automatisierte Systeme lösen lassen. Alle vorhandenen CAPTCHA-Lösungen bieten somit prinzipiell nur Schutz gegen automatisierte Angreifer. Eine typtische Fragestellung ist das Erkennen verfremdeter Schriftzeichen. Diese am häufigsten auftretenden Vertreter der CAPTCHAs sind die Bild-CAPTCHAs [KZ09] [EDHS07]. Durch den Einsatz der Bild-CAPTCHAs auf Seiten wie Facebook, Google oder eBay gelten diese heute als De-facto-Standard. Die Sicherheitsannahme ist, dass menschliche Nutzer diese Verfremdung erkennen und auflösen können, Texterkennungsalgorithmen daran jedoch scheitern. Für das Design von CAPTCHAs ist daher immer eine Abwägung zwischen Sicherheit (vor Angriffen) und Erkennbarkeit (durch Menschen)

*in: Sicherheit 2012. Sicherheit, Schutz und Zuverlässigkeit. Beiträge der 6. Jahrestagung des Fachbereichs Sicherheit der Gesellschaft für Informatik e.V. (GI), Lecture Notes in Informatics (P-195), Michael Waidner, Suri Neeraj (Hrsg.), Köllen-Verlag, Bonn 2012, 199-208.

zu treffen. Bei Auftritt der ersten CAPTCHAs Mitte der 1990er Jahre [Nao96] waren angreifende Programme noch wenig leistungsfähig, so dass es bei auch für Menschen einfachen Fragestellungen bleiben konnte. Im Laufe der Zeit wurden allerdings die automatisierten Angriffe immer ausgefeilter. Durch diese Fortschritte wurde es notwendig, die Schwierigkeit des zu lösenden Problems zu erhöhen. Aktuell ist ein Niveau erreicht, das Menschen bereits vor große Herausforderungen stellt. Der Extremfall unterstreicht die Notwendigkeit guter Usability von CAPTCHAs: Wenn ein Nutzer ein zur Anmeldung verpflichtendes CAPTCHA nicht zu lösen vermag, kann er sein eigentliches Ziel der Interaktion nicht erfüllen - etwa die Benutzung einer Webseite. Es ist somit ein fataler Benutzbarkeitsfehler aufgetreten. CAPTCHAs unterschiedlicher Modi, etwa Bild-, Ton oder Texterkennung, werden in dieser Studie auf ihre Usability getestet. Wir stellen zudem einen Kriterienkatalog vor, der über die reine Erkennungsrate hinausgeht, welche bereits in anderen Studien (siehe Abschnitt 2) untersucht wurde. Diese Untersuchung betrachtet Aspekte der Mensch-Maschine-Interaktion bei der Benutzung von CAPTCHAs und beantwortet folgende Fragen:

- Wie lassen sich Usability-Kriterien für CAPTCHA-Systeme formalisieren?
- Wie gebrauchstauglich sind einzelne CAPTCHAs laut empirischem Nutzertest?
- Welche CAPTCHAs sind im konkreten Anwendungsbereich zu bevorzugen?

2 Sicherheit und Benutzbarkeit von CAPTCHA-Systemen

Aussagen zur Sicherheit einzelner CAPTCHA-Systeme vor Angreifern stehen nicht im Fokus einer Usability-Studie. Um die Auswirkungen der Benutzbarkeitsstudie auf die Realwelt im Gesamtkontext zu betrachten, stellen wir an dieser Stelle dennoch Einschätzungen über die Sicherheit von CAPTCHAs in der Forschung vor. Ein CAPTCHA, das sich in der Benutzbarkeit überlegen zeigt, aber leicht überwunden werden kann, ist für die Lösung der ursprünglichen Aufgabe ebenso ungeeignet wie ein extrem sicheres CAPTCHA, welches sich vom menschlichen Benutzer nur sehr schwer lösen lässt.

Automatisierte Angriffe setzen auf OCR-Systeme (Texterkennungssysteme), Spracherkennungssysteme oder umfangreiche Bibliotheken bekannter CAPTCHA-Lösungen. Bilge et al. [BSBK09] untersuchten die CAPTCHA-Systeme bekannter sozialer Netzwerke und konnten bei ReCAPTCHA (einem CAPTCHA-System von Google) in dessen Bild-CAPTCHA-Version noch Erkennungsraten von 4-7% erreichen. Unter der Annahme, dass pro Tag mehrere hundert Angriffsversuche gestartet werden können, gehen sie aus Angreifersicht dennoch von einem Erfolg aus. Wilkins [Wil09] benutze OCR-Techniken und Bilderkennungsverfahren auf ReCAPTCHA und konnte eine Erkennungsrate von 17,5% erzielen. Ahmad et al. [AYT11] waren in der Lage, 24,7% der Bild-CAPTCHAS von ReCAPTCHA durch fortgeschrittene Segmentation-Erkennungsmethoden zu lösen. Darüber hinaus entwickelten Bursztein et al. [BMM11] eine Möglichkeit, einzelne Bild-CAPTCHA-Systeme mit einer Erfolgsrate zwischen 10% und bis zu 50% zu lösen. Tam

et al. [TSHVA09] konnten in einem Testsample von Audio-CAPTCHAs eine Erkennungsrate von 71% erreichen - sie vergleichen das mit der Rate, welche üblicherweise auch von Menschen bei Audio-CAPTCHAs erreicht wird. Philippe Golle [Gol08] konnte mit Hilfe von Klassifizierungsalgorithmen das Quiz-CAPTCHA Asirra in 10,3% der Versuche überwinden. Hernandez-Castro und Ribagorda [HCR10] untersuchten fortgeschrittene Math-CAPTCHAs und konnten CAPTCHAs selbst aus komplexen mathematischen Aufgaben in 35% der Fälle lösen lassen. Einfache Math-CAPTCHAs, welche keinen OCR-Einsatz, einfache Arithmetik und eine Lösung im niedrigen zweistelligen Zahlenbereich erwarten, sind daher generell als unsicher anzusehen. Aussagen zur Sicherheit von CAPTCHAs sind aber auch immer mit der entsprechenden Lösungsfähigkeit durch Menschen zu vergleichen. Bursztein et al. [BBF+10] geben eine Aussage über die Erfolgsraten menschlicher Nutzer bei verschiedenen CAPTCHA-Implementierungen. Durch einen „Mechanical Turk"-Dienst ließen sie 5.000 CAPTCHAs aus 13 verschiedenen Varianten lösen und kamen im Schnitt auf Lösungsraten von 71% bei Bild-CAPTCHAs und 31,2% bei Audio-CAPTCHAs. Sie schließen daraus, dass die meisten CAPTCHAs für Menschen bereits schwerer zu lösen sind als nötig. Ahmad et al. [AYT11] benennen als Richtwert für die Entwicklung neuer CAPTCHA-Varianten eine Erkennungsrate von mindestens 90% durch Menschen bei maximal 0,01% Erfolg durch automatische Angriffsversuche.

Es kann also keine der unterschiedlichen CAPTCHA-Lösungen einen gegenüber anderen Versionen deutlich überlegenen Schutz bieten. Asirra zeigt sich in der Forschung robuster gegen automatisierte Angriffe als die anderen Varianten, gehört aber auch zu den bis jetzt weniger häufig betrachteten Technologien. Bei mehr als 10% erfolgreicher Angriffe bei allen CAPTCHA-Varianten kann man von keinem effektiven Schutz gegen massierte automatische Lösungsversuche sprechen.

Im Feld der Usabilityforschung zu CAPTCHA-Systemen untersuchten Chellapilla et al. [CS04] Methoden zur Erhöhung des Schwierigkeitsgrades für die OCR-Erkennung bei Bild-CAPTCHAs und deren Auswirkungen auf die Erkennungsleistung von Menschen. Es zeigte sich sowohl beim Einbinden von Störddaten als auch dem Verzerren des Textes ab einem bestimmten Grad ein Einbrechen der Erkennungsleistung von Testpersonen. Baird et al. [BMW05] testeten ebenfalls die Benutzbarkeit von CAPTCHAs beim gleichzeitigen Einsatz sowohl stark wie auch schwach gestreuter Buchstaben. Nach den Bild-CAPTCHAs werden Audio-CAPTCHAs, also das Erkennen von Worten aus Toneinspielungen, laut Yahn und Ahmand [YA08] nächsthäufig eingesetzt. Diese haben theoretisch Erkennungvorteile bei Personen mit beeinträchtigtem Sehvermögen. Bigham und Cavender [BC09] zeigten jedoch, dass neben den Verständnisschwierigkeiten bei den Audio-CAPTCHAs an sich auch ein Benutzbarkeitsproblem hinsichtlich der Player-Schaltflächen besteht. Abseits der Bild- und Audio-CAPTCHAs gibt es noch Nutzerstudien zu Microsoft Asirra, welche auf die menschliche Fähigkeit der Bilder-Erkennung und -kategorisierung abzielt. Hier wurde die Erkennungsrate mit 83,4% angegeben [EDHS07]. Kurt Alfred Kluever [Klu08] betrachtete Video-CAPTCHAs, welche durch den Einsatz multimodaler Aspekte Vorteile versprechen: Es müssen gleichzeitig Bilder und Geräusche vom Benutzer verarbeitet werden. Die Erfolgsrate konnte hier mit 90% angegeben werden.

Effektivität	Wie hoch ist die Erkennungsrate eines CAPTCHAs? Wie viele Versuche benötigt ein Benutzer im Durchschnitt, um ein CAPTCHA zu lösen?
Effizienz	Wie lange braucht ein Benutzer für eine richtige Lösung im Durchschnitt? Kann das CAPTCHA in weniger als 30 Sekunden gelöst werden? [RL03]
Erlernbarkeit	Erkennt ein Benutzer auf den ersten Blick, wie das CAPTCHA korrekt benutzt wird?
Einprägsamkeit	Wie gut kann sich ein Benutzer an ein CAPTCHA-Konzept erinnern?
Zufriedenheit	Wie schwierig findet ein Benutzer ein CAPTCHA? Fühlen sich die Benutzer beim Benutzen eines CAPTCHAs wohl und sind sie gewollt ein bestimmtes System zu benutzen? Welches qualitative Feedback geben die Nutzer?

Abbildung 1: Formalisierte Benutzbarkeitskriterien

3 Benutzbarkeitskriterien für CAPTCHAs

Bislang sind für die Usability von CAPTCHA-Systemen noch keine umfassenden formalen Kriterien für die Gebrauchstauglichkeit von CAPTCHAs definiert. Wir stellen einen erweiterten Kriterienkatalog der Benutzbarkeit von CAPTCHA-Systemen vor. In dieser Studie interpretieren wir diese Kriterien auf Ebene der jeweiligen CAPTCHA-Anwendung. Usabilityprobleme, die durch die jeweils unterschiedliche Implementierung von CAPTCHAs in einzelnen Programmen oder Webseiten entstehen, werden daher nicht berücksichtigt. Der reduzierte Interaktionsumfang von CAPTCHAs auf dieser Abstraktionsebene spricht für die Verwendung der „Anforderungen an die Gebrauchstauglichkeit“ nach ISO-Norm 9241-11[ISO98] (Effektivität, Effizienz und Zufriedenheit), erweitert um Jakob Nielsens qualitative Komponenten [Nie93] (Erlernbarkeit, Einprägsamkeit und Fehlervermeidung) als Basis der Formalisierung. Das Kriterium „Fehlervermeidung“ interpretieren wir in diesem Kriterienkatalog als dem Lösungsproblem inherent, und dadurch nicht als Aspekt der Benutzbarkeit von CAPTCHA-Systemen. Fehler bei der Benutzung von CAPTCHAs sind (auch aufgrund der reduzierten Interaktionsmöglichkeiten) auf die Schwierigkeit der zu lösenden CAPTCHA-Problemstellung zurückzuführen. Eine Vereinfachung des CAPTCHA-Problems zur Verbesserung der Gebrauchstauglichkeit wirkt sich negativ auf die Sicherheit vor automatisierten Angriffen aus, vor denen das CAPTCHA letztendlich schützen soll. Die quantitative Ausprägung der „Fehlervermeidung“ betrachten wir (in Form der Erkennungsrate) in diesem speziellen Kontext als Teilaspekt des Benutzbarkeitskriteriums „Effektivität“.

Es ergeben sich somit als erweiterten Kriterienkatalog der Benutzbarkeit von CAPTCHA-Systemen die in Abbildung 1 formalisierten Usabilityaspekte.

Gesamt	
Merkmal	**Anzahl**
Student	25
Andere	25
Alter	**Anzahl**
14-19	6
20-29	30
30-39	2
40-49	10
50-59	2
ab 60	0
Internetnutzung	**Anzahl**
wenig	12
normal	26
oft	12

Gruppe Andere	
Alter	**Anzahl**
14-19	5
20-29	6
30-39	2
40-49	10
50-59	2
ab 60	0
Internetnutzung	**Anzahl**
wenig	12
normal	11
oft	2

Gruppe Studenten	
Alter	**Anzahl**
14-19	1
20-29	24
30-39	0
40-49	0
50-59	0
ab 60	0
Internetnutzung	**Anzahl**
wenig	0
normal	15
oft	10

Abbildung 2: Demografische Verteilung der Probanden

4 Methodik

Der Benutzertest umfasst 50 Testpersonen. Diese Personengruppe kann in zwei Gruppen unterteilt werden, eine Gruppe mit Studenten der Altersklassen 14-19 sowie 20-29 (Gruppe „Studenten“) und eine Gruppe Nichtstudierender mit einer gemischten Altersstruktur (Gruppe „Andere“). Eine Übersicht über die demografische Verteilung der Probanden gibt Abbildung 2.

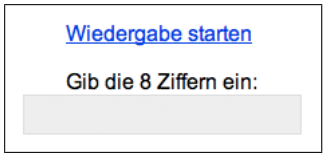
Alle Teilnehmer waren mit den Grundlagen der PC-Bedienung, sowie den gängigen Eingabegeräten Maus und Tastatur vertraut. Die Probanden wurden im Rahmen der Möglichkeiten zufällig ausgewählt. Alle Teilnehmer nahmen freiwillig und ohne (eventuell verzerrendes) Anreizsystem am Test teil. Der Test wurde in einer klassischen Laborumgebung durchgeführt. Dies bedeutet eine geräuscharme wie optisch ruhige Zone, in der sich die Probanden vollständig auf die zu untersuchende Aufgabe konzentrieren konnten. Der Test beginnt mit der Bearbeitung eines Eingangsfragebogens. Danach werden in drei Runden zufällig je ein CAPTCHA-System dem Probanden zur Lösung vorgelegt. Abschließend erfolgt die Erfassung des Feedbacks des Teilnehmers.

Als quantitative Attribute ergeben sich somit die korrekte Lösung des CAPTCHAs (ja/nein), die benötigte Bearbeitungszeit, Notwendigkeit der Hilfefunktion (ja/nein), die jeweilige Runde des Testablaufs (1-3) und die wahrgenommene Schwierigkeit des CAPTCHAs (fünfstufige Likert-Skala).

Unter den möglichen CAPTCHAs musste eine Auswahl getroffen werden, da nicht alle der Systeme die Anforderungen für diesen Benutzertest erfüllen. So wurden rein englischsprachige Systeme oder solche, die nicht frei verfügbar waren, aus praktischen Gründen nicht berücksichtigt. Dennoch konnte eine ausreichende Heterogenität in den Interaktionsmodi der verschiedenen Systeme gewährleistet bleiben. CaptchaAd [capb] ist eine Implementierung, die kurze Videoclips anzeigt, welche Werbung enthalten. Der menschliche Nutzer kann während und nach Betrachten des Videos eine hierzu passende Frage

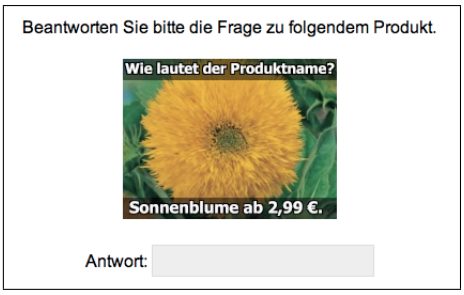
beantworten, z.B. „Wie hoch ist der Preis des Produktes?“. Google reCAPTCHA Audio (im folgenden „Audio-CAPTCHA“ genannt) und reCAPTCHA Bild (im folgenden „Bild-CAPTCHA“) [capd] sind hiermit verglichen klassische Ansätze: Im Audio-CAPTCHA werden acht Zahlen vorgelesen, die der Nutzer dann über die Tastatur eingibt, im Bild-CAPTCHA müssen verzerrt dargestellte Wörter korrekt erkannt und eingegeben werden. Quiz-CAPTCHAs sind in mehreren Varianten im Einsatz, welche auf unterschiedliche Lösungsstrategien abzielen: Math-CAPTCHAs [capc] erfordern die intellektuelle Lösung einer mathematischen Aufgabe, etwa „Was ist die Lösung aus 8 + 9?“. Microsoft Asirra [capa] hingegen setzt auf die Fähigkeit des Menschen, verschiedene Tiere auf Fotos zu unterscheiden, in dem Fall Hunde von Katzen - ein Problem, das für den Menschen einfach zu lösen sein sollte, aber ursprünglich schwer automatisierbar war [EDHS07]. In Abbildung 3 findet sich eine Darstellung der für den Nutzertest herangezogenen CAPTCHA-Systeme.

(a) Microsoft Asirra

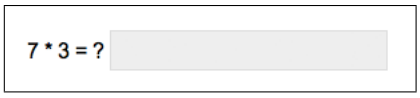


(b) Audio-CAPTCHA

(c) Bild-CAPTCHA



(d) CaptchaAd



(e) Math-CAPTCHA

Abbildung 3: Im Nutzertest untersuchte CAPTCHA-Systeme

5 Auswertung

Die quantifizierbare Messgröße für Angaben zur *Effektivität* ist die Erkennungsrate. Sie ist definiert als der Anteil positiver Lösungen unter allen Versuchen. Das Math-CAPTCHA wurde bei einer Erkennungsrate von 98,67% am besten erkannt. Dem folgte das Bild-CAPTCHA mit 92% Erkenungsrate. Die anderen Systeme Asirra, CaptchaAd und das Audio-CAPTCHA erreichen 84%, 74% respektive 68,87% in der gesamten Testgruppe. Diese Angaben basieren auf 150 Einzelbeobachtungen je CAPTCHA-Variante. Die Aufschlüsselung der Erkennungsrate zeigt Unterschiede zwischen den Runden sowohl im Gesamtbild als auch im Vergleich der beiden Teilnehmergruppen. Am wenigsten schwankt dabei Math-CAPTCHA, das bei der Gruppe „Studenten“ mit einer Erkennungsrate von 100% in allen Runden gar keine Veränderung aufweist. Bei der Gruppe „Andere“ wird Math-CAPTCHA mit 96% in Runde 1 und 2 sowie 100% in Runde 3 fast so gut erkannt wie in der Gruppe Studenten. Beim Audio-CAPTCHA wird bei einem normierten Korrelationskoeffizienten von 0,31 und einem p-value von 0,02 beim Chi-Quadrat-Test ein leichter Zusammenhang zwischen der Runde und dem entsprechenden Ergebnis erkennbar. Mit anderen Worten ist das Audio-CAPTCHA die einzige Lösung, die die Anforderung der „Lernförderlichkeit“ im Sinne der Softwareergonomie erfüllt: Je häufiger das CAPTCHA verwendet wird, desto besser werden die Erkennungsraten. Auffällig bei der Untersuchung von Microsoft Asirra ist der Unterschied in der Erkennungsrate zwischen der Altersgruppe 50-59 Jahre und den restlichen Altersgruppen. Während die Erkennungsraten zwischen 80% bei der Gruppe 14-19 und 90% bei der Gruppe 40-49 schwankt, wurde Microsoft Asirra von den Teilnehmern der Altersgruppe 50-59 Jahre lediglich in 33% der CAPTCHA-Tests korrekt gelöst.

Auch bei der Betrachtung der *Effizienz* der CAPTCHA-Lösung zeigt sich das Math-CAPTCHA als das vorteilhafteste der untersuchten Systeme: Es wurde im Mittel innerhalb von 7,27 Sekunden richtig gelöst. Für eine richtige Lösung beim Bild-CAPTCHA benötigten die Teilnehmer im Durchschnitt mit 17,63 Sekunden knapp 10 Sekunden länger. Es folgen Microsoft Asirra (25,90 Sekunden), CaptchaAd (29,48 Sekunden) und das Audio-CAPTCHA mit 36,18 Sekunden. Bei den Systemen Math-CAPTCHA und CaptchaAd ist in der Bearbeitungsdauer ein Unterschied zwischen korrekter und falscher Lösung erkennbar. Dieser wird in den beiden Fällen durch einen p-value von 0,00 beim Math-CAPTCHA und 0,01 bei CaptchaAd gestützt. Der Bravis-Pearson-Korrelationskoeffizient im Falle CaptchaAd beträgt -0,21. Die Teilnehmer benötigten demnach für einen Fehlversuch länger als für eine korrekte Lösung. Diese Aussage ist auch für das Math-CAPTCHA zutreffend. Durch den niedrigen p-value ist hierbei aber ohnehin von einer Abhängigkeit auszugehen. Die Korrelation ist dabei mit einem Korrelationskoeffizienten von -0,35 ebenfalls stärker als bei CaptchaAd.

Ein Maß für die *Erlernbarkeit* von CAPTCHAs ist der Rate der Verwendung der Hilfefunktionen in den jeweiligen Systemen. Geordnet nach der Häufigkeit des Aufrufens der Hilfefunktion ergibt sich ein anderes Bild als bei der Erkennungsrate. Während das Math-CAPTCHA die höchste Erkennungsrate aufweist, wird hier auch die Hilfe in 6% der Math-CAPTCHA-Tests in Anspruch genommen. Vor dem Math-CAPTCHA liegt noch CaptchaAd mit 11,33% und Microsoft Asirra mit 6,67%. Beim Audio-CAPTCHA wur-

de in 4,67% der Tests die Hilfe benutzt und beim Bild-CAPTCHA in 2,67% der Tests. Auch die beiden Gruppen weisen bei der Benutzung der Hilfe Unterschiede auf. Lediglich ein Student hat die Hilfe beim Audio-CAPTCHA in Runde 1 in Anspruch genommen. Ansonsten wird in dieser Gruppe die Hilfe nicht benötigt. Bei der Gruppe „Andere“ wird dementsprechend die Hilfe häufiger in Anspruch genommen, wobei mit zunehmender Rundenzahl die Inanspruchnahme der Hilfefunktion abnimmt. Eine Ausnahme bildet hier das Bild-CAPTCHA, bei dem in der dritten Runde in 8% der Tests die Hilfe aufgerufen wurde. Bei CaptchaAd wurde die Hilfe in 22,67% aller Tests der Gruppe „Andere“ benutzt. Bei den Asirra CAPTCHAs wurde die Hilfe in 13,33% der Tests benutzt. 12% der Tester in der Gruppe „Andere“ nutzten beim Math-CAPTCHA die Hilfe sowie 8% beim Audio-CAPTCHA. Beim Bild-CAPTCHA wurde die Hilfe in der Gruppe „Andere“ noch in 5,33% der Tests benutzt. Aufgrund des höheren Durchschnittsalters in dieser Gruppe liegt die Vermutung nahe, dass mit zunehmenden Alter des Probanden auch die Verwendung der Hilfefunktion zunimmt. Unterstützt wird diese Aussage durch einen niedrigen p-value von 0,00 und einem normierten Kontingenzkoeffizienten von 0,47 zwischen den Merkmalen „Alter“ und „Anzahl der Aufrufe der Hilfefunktion“.

Die *Einprägsamkeit* eines CAPTCHA-Systems zeigt sich in der benötigten Zeit zum Lösen des CAPTCHAs im Zeitverlauf (bzw. nach Anzahl der Runden im Testdurchlauf). Alle untersuchten Methoden weisen eine Abnahme der zur Lösung benötigten Zeit auf. Unterschiede sind vor allem bei Microsoft Asirra, CaptchaAd und Audio-CAPTCHA nachzuweisen. Microsoft Asirra zeigt die höchste (negative) Abhängigkeit zwischen Runde und Lösungszeit (Korrelationskoeffizient von -0,40). Darauf folgt das Audio-CAPTCHA (-0,35) und CaptchaAd (-0,29) und das Math-CAPTCHA (-0,27). Bei Bild-CAPTCHAs zeigt sich der geringste Effekt der Wiederholung des CAPTCHAs auf die Lösungsrate. Möglicherweise sind die Benutzer mit dieser Methode bereits so vertraut, dass das Verfahren an sich keine Auswirkungen mehr auf die Lösungszeit hat.

Die *Zufriedenheit* der Probanden mit dem System wurde in den Fragestellungen nach Ende des Tests erfasst. Das Audio-CAPTCHA wurde mit einem Wert von 2,85 als am schwersten empfunden, gefolgt von CaptchaAd mit einem Wert von 1,85. Das Bild-CAPTCHA wurde mit 1,72 bewertet. Microsoft Asirra folgt mit einem Wert von 1,44 und das Math-CAPTCHA mit 1,11. Abbildung 4 zeigt die Streuung der Bewertungspunkte. Das Math-CAPTCHA wurde in keinem Einzeltest schwerer als 3 (auf einer Skala von 1-5) empfunden. Microsoft Asirra wurde über alle CAPTCHA-Tests nie schwerer als 4 bewertet. Das Bild-CAPTCHA und CaptchaAd wurden im Vergleich zum Audio-CAPTCHA nur in wenigen Fällen als schwer empfunden. Die Bewertungen beim Audio-CAPTCHA sind im Gegensatz zu den anderen Methoden am weitesten gestreut. Außerdem hat das Audio-CAPTCHA auch die höchste Anzahl an schlechten Bewertungen und die wenigsten Probanden bewerteten es als leicht lösbar.

Die freie Feedbackmöglichkeit am Ende des Tests ermöglichte auch qualitative Aussagen zu den einzelnen CAPTCHA-Systemen, die über die rein quantifizierbaren bzw. die strukturiert qualifizierbaren Merkmale hinausgeht: Vier Teilnehmer empfanden das Audio-CAPTCHA zu schwer oder nicht verständlich. Ein weiterer Teilnehmer stufte Audio-CAPTCHAs als zu umständlich und zu fehleranfällig und deshalb nicht praxistauglich ein. Zwei Teilnehmer empfanden die Fragestellungen im CaptchaAd nicht präzise ge-

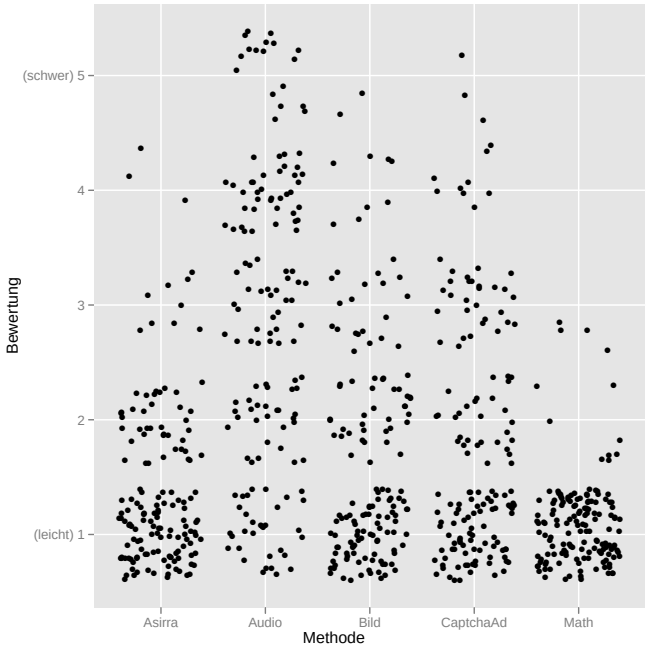


Abbildung 4: Bewertung der CAPTCHA-Systeme durch die Probanden

nug bzw. wussten nicht, was aus den gegebenen Informationen die korrekte Lösung ist. Microsft Asirra wird von zwei Teilnehmern als angenehm und einfach empfunden, ein Teilnehmer hingegen fühlte sich von Asirra genervt. Ein anderer Teilnehmer empfand das Bild-CAPTCHA als zu schwer lesbar. Ein Proband hielt das Math-CAPTCHA für nicht sicher.

Abbildung 5 zeigt abschließend die unterschiedlichen CAPTCHA-Systeme im Hinblick auf die Erfüllung der Usability-Kriterien erweitert um eine qualitative Aussage zur Zugänglichkeit (Accessibility).

6 Schlussbemerkungen

Bei der Betrachtung der untersuchten CAPTCHA-Lösungen bezüglich ihrer Sicherheit zeigte sich bereits, dass keine der Lösungen ihren jeweiligen Alternativen generell überlegen war. Asirra wurde in Studien nur in 10% der Angriffsversuche überwunden, kann allerdings in der Benutzbarkeit nicht überzeugen. Das Math-CAPTCHA wiederum ist hinsichtlich der Kriterien der Benutzbarkeit das System, welches die gestellten Usability-Anforderungen am besten erfüllt. Allerdings bietet es nur Schutz gegen simple Loginversuche. Es zeigt sich am Ende unserer Untersuchungen, dass die be-

	Erlernbarkeit	Effektivität	Effizienz	Einprägsamkeit	Zufriedenheit	Zugänglichkeit
Quiz	✔	✔	✔	-	✔	✔
Bild	✔	✔	✔	✔	✔	O
Asirra	✔	O	✔	-	✔	O
CaptchaAd	O	O	O	-	✔	×
Audio	✔	O	×	-	×	O

Legende:
✔ | Anforderung erfüllt
O | Anforderung teilweise erfüllt
× | Anforderung nicht erfüllt
- | keine Aussage möglich

Abbildung 5: Benutzbarkeitsbewertung von CAPTCHA-Systemen

kannten Bild-CAPTCHAs in der Kombination aus Sicherheit und Benutzbarkeit am deutlichsten überzeugen können. Sie sind nach Asirra das robusteste, und nach den Math-CAPTCHAs das benutzerfreundlichste Verfahren, ohne jedoch die kritischen Einschränkungen (schlechte Benutzbarkeit oder fehlende Sicherheit) dieser beiden zu teilen.

Weitere Forschungsansätze liegen in der Betrachtung anderer CAPTCHA-Systeme. Aktuell sind Benutzer mit den üblichen Bild-CAPTCHAs deutlich vertrauter als mit bereits vorhandenen, aber weniger häufig eingesetzten Alternativen. In einigem zeitlichen Abstand könnte sich diese Verzerrung verändert haben. Es konnten im Laufe der Durchführung des Nutzertests auch Effekte nachgewiesen werden, die nicht sofort zu erklären waren. So ist etwa noch kein Grund für die Korrelation aus dem Alter und der erzielten Erkennungsrate klar zu erkennen. Es mag an der allgemein höheren Computeraffinität jüngerer Menschen liegen. Eine abschließende Untersuchung dieser Hypothese steht aber noch aus.

Es ist aus den beobachteten Fortschritten in den Angriffstechniken auch eine theoretische Grenze für den Einsatz von CAPTCHAs abzusehen. Sobald deren Erkennungsraten im Mittel diejenigen der menschlichen Nutzer übersteigen, muss die Mensch-Computer-Unterscheidung auf andere, noch zu ergründende Arten geschehen. Zur Erhöhung von sowohl Sicherheit als auch Usability sind Fortschritte im Design multimodaler CAPTCHAs ein naheliegender Ansatz. CaptchaAd zeigte im Test Schwächen sowohl in den Benutzbarkeitskriterien, als auch in der Zugänglichkeit. Allerdings steht ein Nachweis der zusätzlichen Sicherheit multimodaler Systeme noch aus.

Unabhängig vom Konzept des inversen Turing Tests sind auch andere Formen des Nachweises legitimer Nutzer möglich. Beispielsweise erfordern oder ermöglichen Google und Facebook die Angabe einer Mobilfunknummer, an die eine Textnachricht mit einem Sicherheitscode gesendet wird. Die Sicherheitsannahme dabei ist, dass der Aufwand zur massenhaften Generierung von valider mobiler Anschlüsse hierbei den erwarteten Nutzen für Spammer übersteigt. Diesen Ansatz greift jedoch noch tiefer in die persönlichen Daten der Nutzer ein. Es bleibt abzuwarten, ob die Sicherheitsbedenken der Anbieter sich gegen den Willen zur Datensparsamkeit der Benutzer durchsetzen.

Literatur

[AYT11] Salah El Ahmad Ahmad, Jeff Yan und Mohamad Tayara. The Robustness of Google CAPTCHAs. Bericht, Newcastle University, 2011.

[BBF+10] Elie Bursztein, Steven Bethard, Celine Fabry, John C. Mitchell und Dan Jurafsky. How Good Are Humans at Solving CAPTCHAs? A Large Scale Evaluation. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, SP '10, Seiten 399–413, Washington, DC, USA, 2010. IEEE Computer Society.

[BC09] Jeffrey P. Bigham und Anna Cavender. Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use. In Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson und Saul Greenberg, Hrsg., *CHI*, Seiten 1829–1838. ACM, 2009.

[BMM11] Elie Bursztein, Matthieu Martin und John C. Mitchell. Text-based CAPTCHA Strengths and Weaknessses. *ACM CSS '11*, 2011.

[BMW05] Henry S. Baird, Michael A. Moll und Sui-Yu Wang. A Highly Legible CAPTCHA That Resists Segmentation Attacks. In Henry S. Baird und Daniel P. Lopresti, Hrsg., *HIP*, Jgg. 3517 of *Lecture Notes in Computer Science*, Seiten 27–41. Springer, 2005.

[BSBK09] Leyla Bilge, Thorsten Strufe, Davide Balzarotti und Engin Kirda. All Your Contacts Are Belong To Us: Automated Identity Theft Attacks on Social Networks. *WWW 2009 Madrid*, 2009.

[capa] ASIRRA. Website. http://research.microsoft.com/en-us/um/redmond/projects/asirra.

[capb] CaptchaAd. Website. http://www.captchaad.com.

[capc] MathCaptcha. Website. https://github.com/niklas/rails-math-captcha.

[capd] reCAPTCHA: Stop Spam, Read Books. Website. http://www.google.com/recaptcha.

[CS04] Kumar Chellapilla und Patrice Y. Simard. Using Machine Learning to Break Visual Human Interaction Proofs (HIPs). In *NIPS*, 2004.

[EDHS07] Jeremy Elson, John R. Douceur, Jon Howell und Jared Saul. Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization. In *CCS '07: Proceedings of the 14th ACM conference on Computer and communications security*, Seiten 366–374. ACM, 2007.

[Gol08] Philippe Golle. Machine learning attacks against the Asirra CAPTCHA. In Peng Ning, Paul F. Syverson und Somesh Jha, Hrsg., *ACM Conference on Computer and Communications Security*, Seiten 535–542. ACM, 2008.

[HCR10] Carlos Javier Hernandez-Castro und Arturo Ribagorda. Pitfalls in CAPTCHA design and implementation: The Math CAPTCHA, a case study. *Computer Security*, (29):141–157, 2010.

[ISO98] Ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: Guidance on usability, 1998. ISO 9241-11:1998 Norm.

[Klu08] Kurt Alfred Kluever. Evaluation the Usability and Security of a Video CAPTCHA. Diplomarbeit, Rochester Institute of Technology, August 2008.

[KZ09] Kurt Alfred Kluever und Richard Zanibbi. Balancing usability and security in a video CAPTCHA. In Lorrie Faith Cranor, Hrsg., *SOUPS*, ACM International Conference Proceeding Series. ACM, 2009.

[Nao96] Moni Naor. Verification of a human in the loop or Identification via the Turing Test. September 1996.

[Nie93] Jakob Nielsen. *Usability Engineering*. Academic Press, San Diego, 1993.

[RL03] Yong Rui und Zicheng Liu. ARTiFACIAL: automated reverse turing test using FACIAL features. In Lawrence A. Rowe, Harrick M. Vin, Thomas Plagemann, Prashant J. Shenoy und John R. Smith, Hrsg., *ACM Multimedia*, Seiten 295–298. ACM, 2003.

[TSHVA09] Jennifer Tam, Jiri Simsa, Sean Hyde und Luis Von Ahn. Breaking audio CAPTCHAs. *Adv. Neu. Inform. Process. Syst.*, 21:1625–1632, 2009.

[Wil09] Jonathan Wilkins. Strong CAPTCHA Guidelines, December 2009. http://frederic.plename/public/documents/captcha.pdf.

[YA08] Jeff Yan und Salah El Ahmad Ahmad. Usability of CAPTCHAs or usability issues in CAPTCHA design. In Lorrie Faith Cranor, Hrsg., *SOUPS*, ACM International Conference Proceeding Series, Seiten 44–52. ACM, 2008.