RUHR-UNIVERSITÄT BOCHUM

RUB

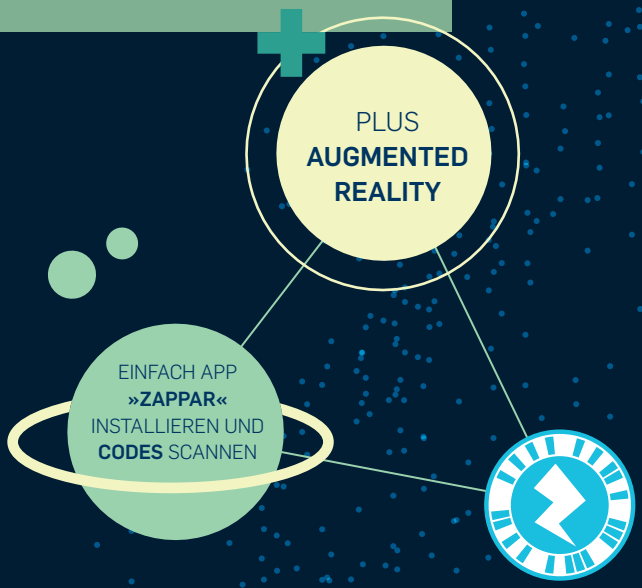
RUBIN

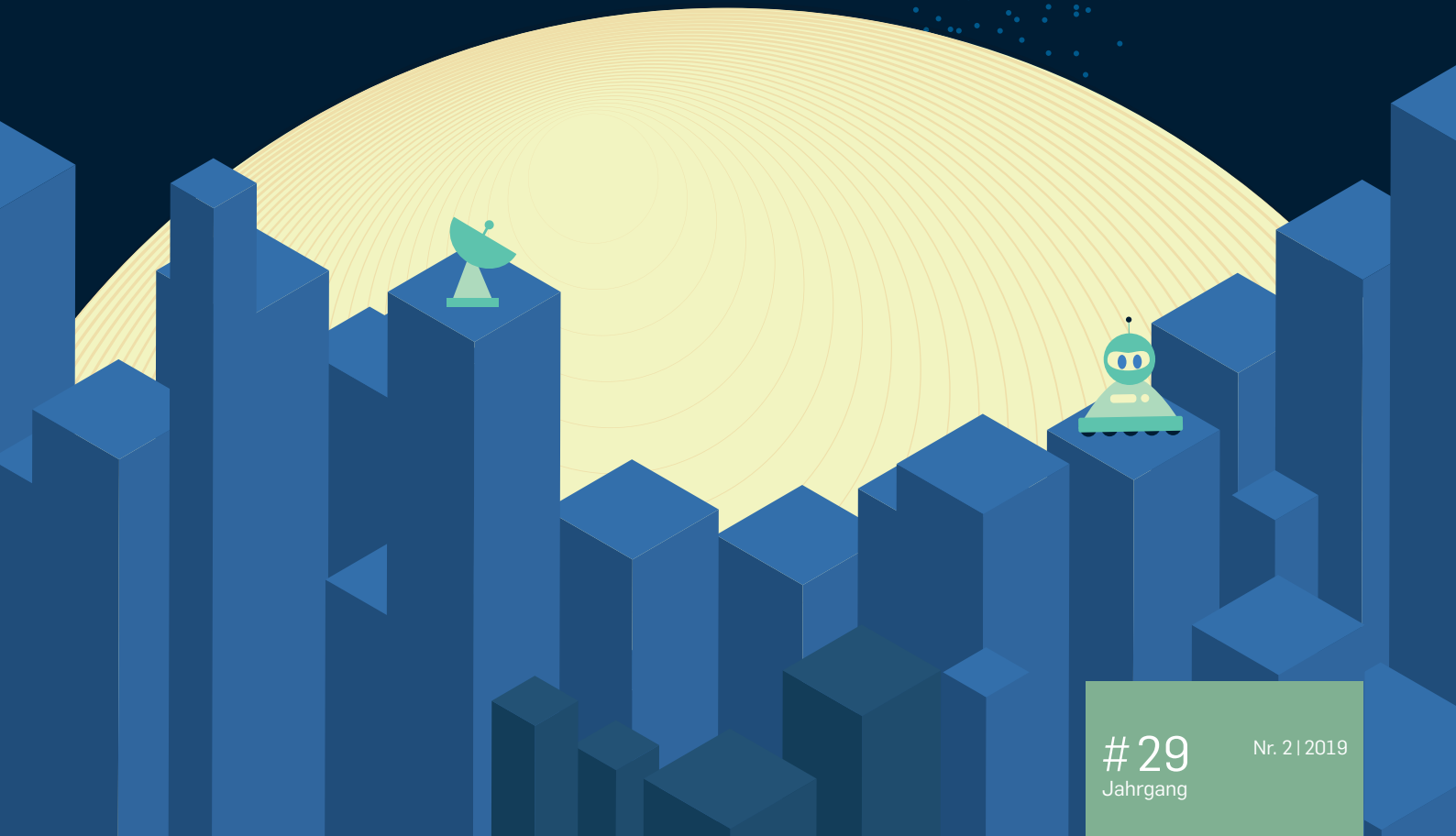
WISSENSCHAFTSMAGAZIN

Schwerpunkt

VIRTUELLE WELTEN

PSYCHISCH KRANKE AVATARE
GEHEIME BOTSCHAFTEN FÜR ALEXA & CO.
KÜNSTLICHE UN-INTELLIGENZ





29
Jahrgang

Nr. 2 | 2019

WIE SPRACH-ASSISTENTEN UNHÖRBARE BEFEHLE BEFOLGEN

Was für einen Menschen nach einem harmlosen Musikstück klingt, kann für eine Maschine die Anweisung sein, eine bestimmte Aktion auszuführen.

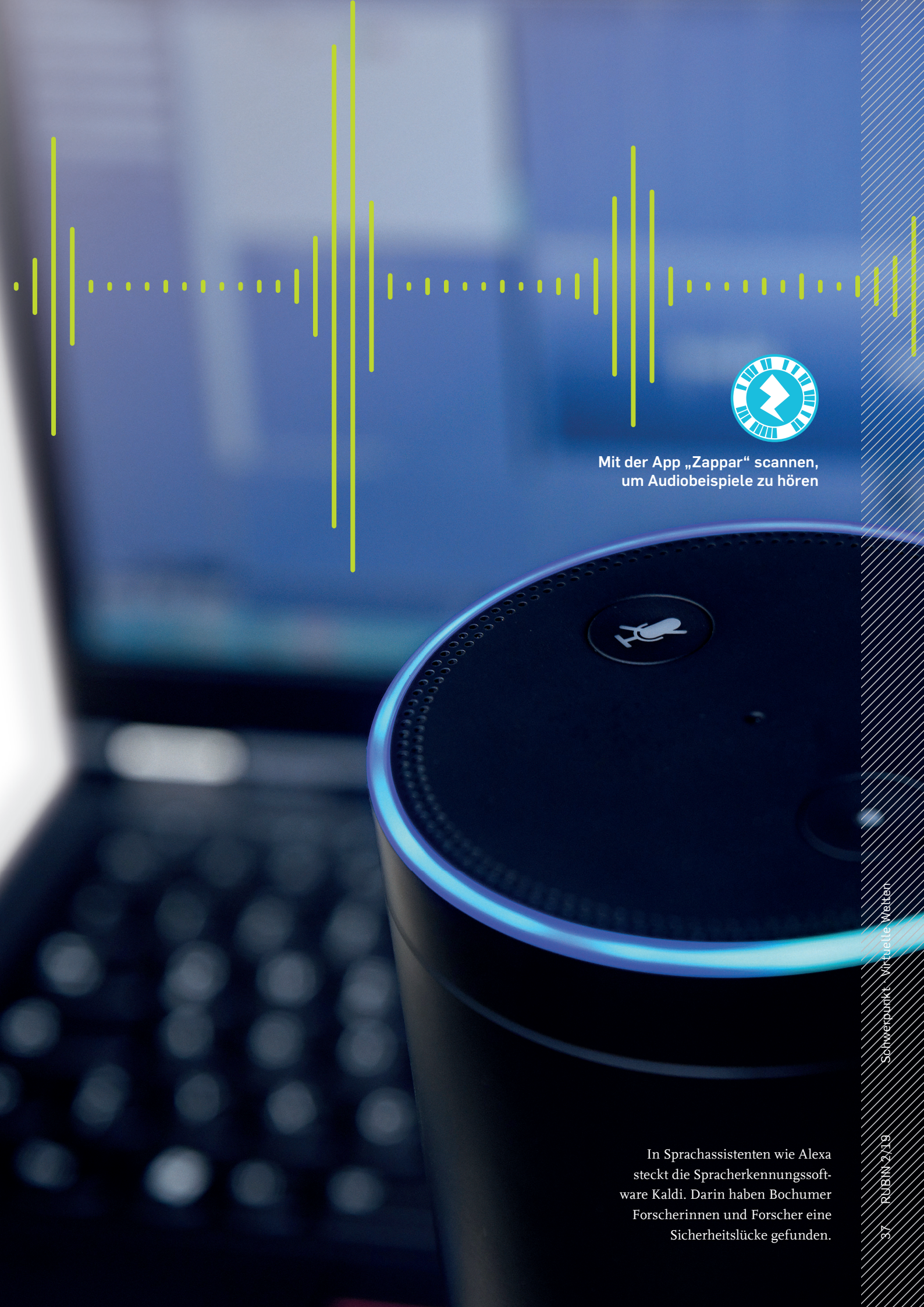
Viel besser als zu den Anfängen der Spracherkennungssysteme verstehen Alexa, Siri und Co. heute, was Menschen ihnen sagen. Manchmal verstehen sie sogar Dinge, die der Mensch nicht hören kann. Eine Sicherheitslücke, wie die IT-Experten vom Bochumer Horst-Görtz-Institut für IT-Sicherheit (HGI) wissen. Ihnen gelang es, beliebige Befehle für Sprachassistenten in unterschiedlichen Arten von Audiosignalen zu verstecken, zum Beispiel in Musik, Sprache oder Vogelgezwitscher. Solange diese Angriffe nur der Forschung dienen, passiert dabei nichts Schlimmes. Ein bösartiger Angreifer könnte auf diese Weise aber beispielsweise einen Song, der im Radio abgespielt wird, so manipulieren, dass er den Befehl enthält, ein bestimmtes Produkt zu kaufen oder Kontrolle über ein sprachgesteuertes Smart Home zu übernehmen.

In der Fachsprache werden solche Angriffe als Adversarial Examples bezeichnet. Lea Schönherr aus der HGI-Arbeitsgruppe Kognitive Signalverarbeitung entwickelt sie in ihrer Doktorarbeit im Team von Prof. Dr. Dorothea Kolossa. „Wir nutzen dafür das psychoakustische Modell des Hörens“, erzählt Lea Schönherr. Wenn das Gehör damit beschäftigt ist, einen Ton einer bestimmten Frequenz zu verarbeiten, können Menschen für einige Millisekunden andere leisere Töne nicht mehr wahrnehmen. Genau in diesen Bereichen verstecken die Forscherinnen und Forscher die geheimen Befehle für die Maschinen. Für den Menschen klingt die zusätzliche Information wie zufälliges Rauschen, das im Gesamtsignal kaum oder gar nicht auffällt. Für den Sprachassistenten ändert es jedoch den Sinn: Der Mensch hört Aussage A, während die Maschine Aussage B versteht.

Ihre Angriffe testete Lea Schönherr an dem Spracherkennungssystem Kaldi, einem Open-Source-System, welches in Amazons Alexa und vielen anderen Sprachassistenten enthalten ist. Sie versteckte unhörbare Befehle in unterschiedlichen Audiosignalen und überprüfte, welche Information Kaldi daraus decodierte. Tatsächlich verstand das Spracherkennungssystem die geheimen Befehle zuverlässig.

Zunächst funktionierte dieser Angriff nicht über den Luftweg, sondern nur, wenn Lea Schönherr die manipulierten Audiodateien direkt in Kaldi hineinspielte. Mittlerweile kommen die geheimen Botschaften aber auch an, wenn die Forscherin dem Spracherkennungssystem die Audiosignale über einen Lautsprecher vorspielt. „Das ist viel komplizierter“, erklärt sie. „Denn der Raum, in dem die Datei abgespielt wird, beeinflusst den Klang.“ Ein Musikstück hört sich etwa anders an, wenn es in einem Kino ertönt, als wenn es über die Lautsprecherboxen eines Autos gespielt wird. Die Größe des Raums, das Material der Wände und die Position des Lautsprechers im Raum spielen dabei eine Rolle.

All diese Parameter muss Lea Schönherr berücksichtigen, wenn sie eine Audiodatei erzeugen will, die ein Sprachassistent in einem bestimmten Raum verstehen können soll. Dabei hilft die sogenannte Raumimpulsantwort. Sie beschreibt, wie ein Raum den Schall reflektiert und so den Klang ▸



Mit der App „Zappar“ scannen, um Audiobeispiele zu hören

In Sprachassistenten wie Alexa steckt die Spracherkennungssoftware Kaldi. Darin haben Bochumer Forscherinnen und Forscher eine Sicherheitslücke gefunden.

Die Forscherinnen und Forscher manipulieren Audiodateien so, dass Maschinen eine ganze andere Aussage verstehen als Menschen.

verändert. „Wenn wir wissen, in welchem Raum der Angriff erfolgen soll, können wir die Raumimpulsantwort mit speziellen Computerprogrammen simulieren und beim Erzeugen der manipulierten Audiodatei berücksichtigen“, erklärt Lea Schönherr. Dass das funktioniert, hat die Forscherin bereits gezeigt. Im Testraum an der RUB decodierte Kaldi wie gewünscht die geheimen Botschaften, die die Forscherin zuvor in verschiedenen Tonsignalen versteckt hatte.

Gegenmaßnahmen entwickeln

„Wir können den Angriff also für einen bestimmten Raum maßschneidern“, berichtet die Kommunikationstechnikerin. „Kürzlich ist es uns aber sogar gelungen, einen allgemeinen Angriff durchzuführen, der keine Vorinformationen über den Raum benötigt, und trotzdem genauso gut oder sogar noch besser auf dem Luftweg funktioniert.“

Künftig plant die Wissenschaftlerin auch Tests mit auf dem Markt erhältlichen Sprachassistenten. Da Sprachassistenten aktuell nicht in sicherheitskritischen Bereichen im Einsatz sind, sondern lediglich dem Komfort dienen, können die Adversarial Examples derzeit keinen großen Schaden anrichten. Daher sei es noch früh genug, die Sicherheitslücke zu schließen, meinen die Forscher am Bochumer HGI. Im Exzellenzcluster Casa, kurz für Cyber Security in the Age of Large-Scale Adversaries, kooperiert die Arbeitsgruppe Kognitive Signalverarbeitung, die die Angriffe entwickelt hat, mit dem Lehrstuhl für Systemsicherheit von Prof. Dr. Thorsten Holz, dessen Team an Gegenmaßnahmen dazu arbeitet. Die IT-Sicherheitsforscher wollen Kaldi beibringen, für Menschen nicht hörbare Bereiche in Audiosignalen auszusortieren und nur das zu hören, was übrig bleibt. „Im Grunde soll die Erkennung der Maschine mehr wie das menschliche Gehör funktionieren, sodass es schwieriger wird, geheime Botschaften in Audiodateien zu verstecken“, erklärt Thorsten Eisenhofer, der in seiner Promotion die Sicherheit von intelligenten Systemen untersucht.

Die Forscher können zwar nicht verhindern, dass Angreifer Audiodateien manipulieren. Wenn diese Manipulation aber in den für Menschen hörbaren Bereichen platziert werden müsste, weil die Spracherkenner den Rest aussortieren, so ließen sich die Angriffe nicht so leicht verstecken. „Wir wollen also, dass der Mensch wenigstens hören kann, wenn mit einer Audiodatei etwas nicht stimmt“, so der Forscher. „Im besten Fall muss ein Angreifer die Audiodatei so weit manipulieren, dass diese mehr wie die verstecke Botschaft klingt als wie das eigentlich Gesagte.“ Die Idee: Wenn der Spracherkenner die für Menschen nicht hörbaren Bereiche eines Audiosignals aussortiert, müsste ein Angreifer auf die hörbaren Bereiche ausweichen, um seine Befehle zu platzieren. Um das zu realisieren, nutzt Thorsten Eisenhofer das MP3-Prinzip.

MP3-Dateien werden komprimiert, indem für Menschen nicht hörbare Bereiche gelöscht werden – genau das ist es, was

In beliebigen Audiodateien wie Sprache, Musik oder Umgebungsgeräuschen – zum Beispiel Vogelgezwitscher – kann das Team geheime Botschaften für die Sprachassistenten verstecken. Früher funktionierten die Angriffe nur, wenn die manipulierten Dateien als Daten in die Spracherkennungssoftware gefüttert wurden. Heute gelingen sie auch, wenn die Audiodateien über Lautsprecher abgespielt werden.

Lea Schönherr und Thorsten Eisenhofer promovieren am Horst-Görtz-Institut für IT-Sicherheit.

die Verteidigungsstrategie gegen die Adversarial Examples auch vorsieht. Eisenhofer kombinierte Kaldi daher mit einem MP3-Encoder, der die Audiodateien zunächst bereinigt, bevor sie zum eigentlichen Spracherkenner gelangen. Die Tests ergaben, dass Kaldi die geheimen Botschaften tatsächlich nicht mehr verstand, es sei denn sie wurden in die für Menschen wahrnehmbaren Bereiche verschoben. „Das veränderte die Audiodatei aber merklich“, berichtet Thorsten Eisenhofer. „Die Störgeräusche, in denen die geheimen Befehle versteckt sind, wurden deutlich hörbar.“

Gleichzeitig blieb Kaldis Spracherkennungsperformance trotz der MP3-Bereinigung vergleichbar gut wie die Spracherkennung für nicht bereinigte Dateien. Allerdings nur, wenn das System auch mit MP3-komprimierten Dateien trainiert wurde. „In Kaldi arbeitet ein Machine-Learning-Modell“, erklärt Thorsten Eisenhofer diesen Umstand. Dieses Modell ist sozusagen eine künstliche Intelligenz, die mithilfe vieler Audiodateien als Lernmaterial trainiert wird, den Sinn von Tonsignalen zu interpretieren. Nur wenn Kaldi mit MP3-komprimierten Daten trainiert wird, kann es diese später auch verstehen. Mit diesem Training konnte Thorsten Eisenhofer das Spracherkennungssystem dazu bringen, alles zu verstehen, was es verstehen soll – aber eben nicht mehr.

Text: jwe, Fotos: rs

REDAKTIONSSCHLUSS

Rund 30 Jahre lang war Helga Schulze als wissenschaftliche Zeichnerin an der Medizinischen Fakultät der RUB tätig und hat anatomische Abbildungen angefertigt. Privat engagiert sich die Diplom-Biologin für den Artenschutz und betreibt unter anderem eine Rettungsstation für Loris. Die Halbaffen werden immer wieder verbotenerweise als Haustiere gehandelt, weil sie so niedlich aussehen. Das hier gezeigte Motiv hat Helga Schulze einem Aberglauben auf Sri Lanka gewidmet, der besagt, Loris würden nachts Pfauen angreifen und erwürgen.

Mehr über die Arbeit von Helga Schulze:
↗ *news.rub.de/wissenschaftlich-zeichnen*

Bild: Helga Schulze

IIMPRESSUM

HERAUSGEBER: Rektorat der Ruhr-Universität Bochum in Verbindung mit dem Dezernat Hochschulkommunikation (Abteilung Wissenschaftskommunikation) der Ruhr-Universität Bochum

WISSENSCHAFTLICHER BEIRAT: Prof. Dr. Gabriele Bellenberg (Philosophie und Erziehungswissenschaften), Prof. Dr. Astrid Deuber-Mankowsky (Philologie), Prof. Dr. Constantin Goschler (Geschichtswissenschaften), Prof. Dr. Markus Kaltenborn (Jura), Prof. Dr. Achim von Keudell (Physik und Astronomie), Prof. Dr. Dorothea Kolossa (Elektrotechnik/Informationstechnik), Prof. Dr. Denise Manahan-Vaughan (Medizin), Prof. Dr. Martin Muhler (Chemie), Prof. Dr. Franz Narberhaus (Biologie), Prof. Dr. Andreas Ostendorf (Prorektor für Forschung, Transfer und wissenschaftlichen Nachwuchs), Prof. Dr. Martin Tegenthoff (Medizin), Prof. Dr. Martin Werding (Sozialwissenschaft), Prof. Dr. Marc Wichern (Bau- und Umweltingenieurwissenschaft), Prof. Dr. Peter Wick (Evangelische Theologie), Prof. Dr. Stefan Winter (Wirtschaftswissenschaft)

REDAKTIONSANSCHRIFT: Dezernat Hochschulkommunikation, Abteilung Wissenschaftskommunikation, Ruhr-Universität Bochum, 44780 Bochum, Tel.: 0234/32-25228, Fax: 0234/32-14136, rubin@rub.de, news.rub.de/rubin

REDAKTION: Dr. Julia Weiler (jwe, Redaktionsleitung); Meike Drießen (md); Raffaela Römer (rr)

FOTOGRAFIE: Damian Gorczany (dg), Hofsteder Str. 66, 44809 Bochum, Tel.: 0176/29706008, damiangorczany@yahoo.de, www.damiangorczany.de; Roberto Schirdewahn (rs), Offerkämpe 5, 48163 Münster, Tel.: 0172/4206216, post@people-fotograf.de, www.wasaufdieaugen.de

COVER: Agentur der RUB

BILDNACHWEISE INHALTSVERZEICHNIS: Teaserfotos für die Seiten 12, 54 und 58: Damian Gorczany; Teaserfotos für die Seiten 32 und 50: Roberto Schirdewahn

GRAFIK, ILLUSTRATION, ANIMATION, LAYOUT UND SATZ: Agentur der RUB, www.rub.de/agentur

DRUCK: VMK Druckerei GmbH, Faberstraße 17, 67590 Monsheim, Tel.: 06243/909-110, www.vmk-druckerei.de

AUFLAGE: 4.700

ANZEIGENVERWALTUNG UND -HERSTELLUNG: VMK GmbH & Co. KG, Faberstraße 17, 67590 Monsheim, Tel.: 06243/909-0, www.vmk-verlag.de

BEZUG: RUBIN erscheint zweimal jährlich und ist erhältlich im Dezernat Hochschulkommunikation (Abteilung Wissenschaftskommunikation) der Ruhr-Universität Bochum. Das Heft kann kostenlos abonniert werden unter news.rub.de/rubin/abo. Das Abonnement kann per E-Mail an rubin@rub.de gekündigt werden.

ISSN: 0942-6639