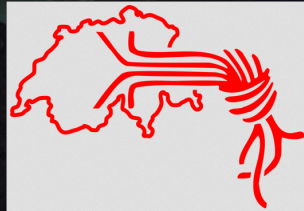


Die Macht der Manipulation: DeepFakes und der Ruf nach Regulierung sogenannter KI

AI-Strategie-Forum (C-Level)

Hernâni Marques (Chaos Computer Club Schweiz)
hernani@ccc-ch.ch // [@vecirex](https://www.vecirex.ch) // [@hernani@chaos.social](https://www.chaos.social) // [vecirex.net](https://www.vecirex.net)

Pfäffikon SZ, 22. Juni 2023



Chaos Computer Club Schweiz? (1)

Generelles

- Dezentrale Hackerorganisation der Zivilgesellschaft
 - Weder Staat noch Privatwirtschaft verpflichtet
 - Selbstzweck: Spass am Gerät
 - Öffentlichkeitsfunktion: Aufklärung
- Seit 2012 Verbandstruktur mit aktuell sieben Kollektivmitgliedern als «Hackerspaces»
- Hackerethik mit Fokus Privatsphäre / Informationsfreiheit / Dezentralisierung / Verantwortungsbewusstsein



Chaos Computer Club Schweiz? (2)

Hackerethik

- Der Zugang zu Computern und allem, was einem zeigen kann, wie diese Welt funktioniert, sollte unbegrenzt und vollständig sein.
- Alle Informationen müssen frei sein.
- Misstrauere Autoritäten – fördere Dezentralisierung.
- Beurteile einen Hacker nach dem, was er tut, und nicht nach üblichen Kriterien wie Aussehen, Alter, Herkunft, Spezies, Geschlecht oder gesellschaftliche Stellung.

• ...



Chaos Computer Club Schweiz? (3)

Hackerethik

- ...
- Man kann mit einem Computer Kunst und Schönheit schaffen.
- Computer können dein Leben zum Besseren verändern.
- Mülle nicht in den Daten anderer Leute.
- Öffentliche Daten nützen, private Daten schützen.



Grundsatz: Mit Digitalisierung umgehen, heisst Ausmass sehen

- Alles wird digitalisiert und vernetzt: Sinn, Nutzen, Risiko, potenzielles Schadensmass?
- Technikfolgenabschätzung fehlt häufig: muss Schaden zuerst eintreten?
- Skalierungseffekte durch Computerisierung bleiben oft unterschätzt: Beispiel E-Voting mit skalierbarem Angriffspotenzial oder Elektronisches Patientendossier mit hochzentralisiertem Datensilo; prinzipiell durch Einzelne ausbeutbar
- Wissen um Funktionsweise von Computern und Internet sowie Bewusstsein für Datenspuren / Sicherheit in aller Regel in der Breite ungeschärft



Fakes per se?

Im Osten wie Westen nichts Neues

- Entfernen von Gesichtern aus Bildern: z. B. Trotzki im Rahmen von UdSSR-Auftritten (durch Retuschieren und Neuschreibung von Geschichte à la Orwells «1984»)
- Erstellen von Fake-Beweisen: z. B. für den Irak-Krieg 2003 (durch unscharfe Bilder und massenmediale Glaubhaftmachung)
- Auch im Rahmen des Ukraine-Krieges sind gerade Bildmaterialien schwer zu beurteilen
- Verbreiten massenhafter Falschbehauptungen in Social-Media, auch in automatisierter Form: vgl. Cambridge-Analytica-Skandal um Facebook



Wesen von «KI»?

- Heute wird auf Basis von (massenhaft) Trainingsdaten erreicht, dass Muster gelernt werden; regelbasierte Systeme mit «Weltwissen» waren wenig erfolgreich
- Konkrete Regeln für das Erkannte / Generierte ist Machine-Learning-System nicht inhärent: von «Intelligenz» im Sinne von Verständnis zu sprechen, ist entsprechend gewagt
- «Gesunder Menschenverstand» ist unvorhanden
- Wo zuwenig Daten bestehen (Sparse-Data-Problem), schlagen «KI»-Systeme völlig fehl; «Denken»: Fehlanzeige
- Ethische Grundsätze / Moral sind nicht inhärent; das Umfeld der konkreten Trainingsdaten oder nachgelagerte Filter bestimmen die Grenzen

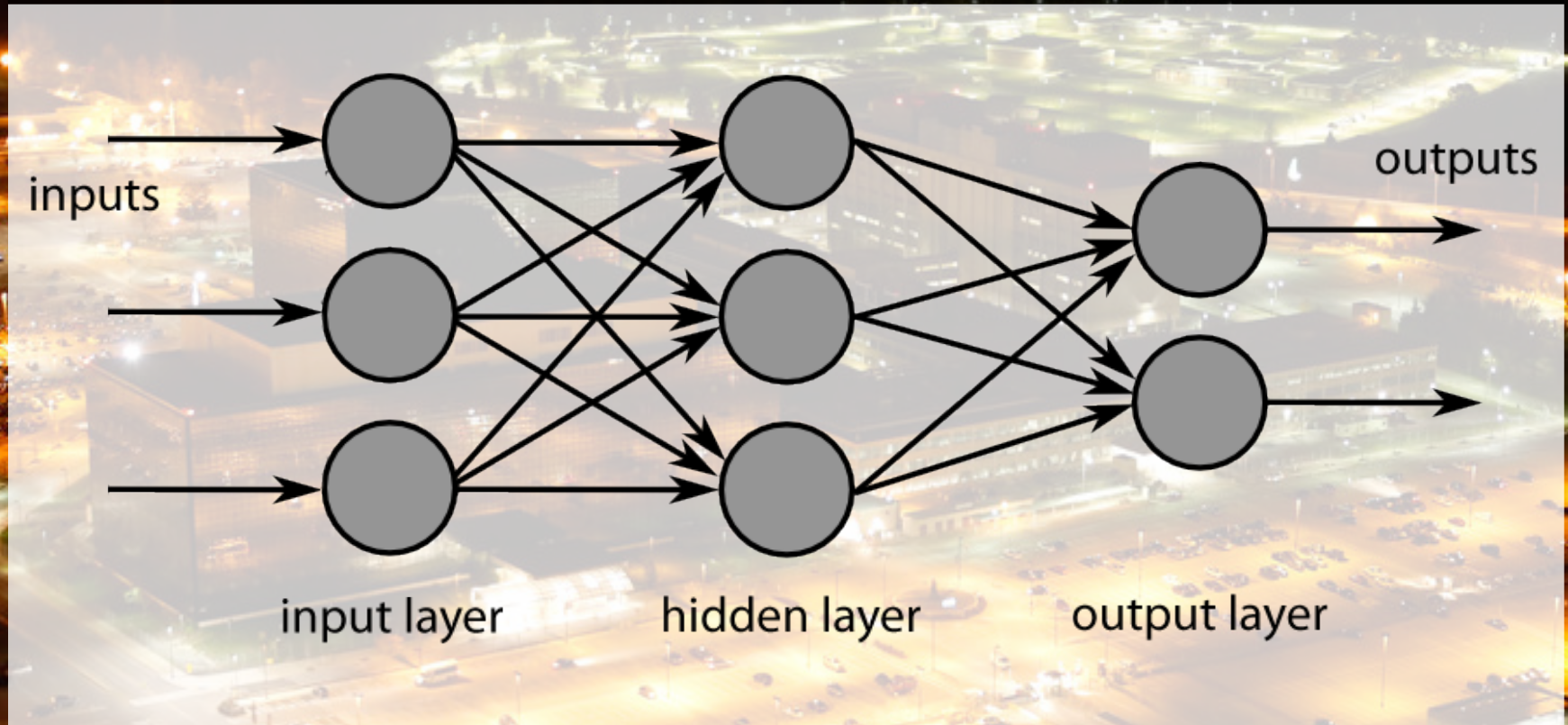


Wesen von DeepFakes? (1)

- Bei DeepFakes sind Machine-Learning-Systeme gemeint, die auf Basis neuronaler Netzwerke Muster erlernen
- Nachgeahmt soll (idealerweise) das menschliche Gehirn werden, mit untereinander verknüpfter «Synapsen»
- Benötigt werden grosse Datenmengen im Sinne von Big-Data; oft weit mehr als ein Mensch (bewusst) weiss, um Ähnliches zu leisten
- Anwendungen sind Generierung / Erkennung von Text, Ton (z. B. Stimmen), Bildern (z. B. Gesichtern) oder Filmen



Wesen von DeepFakes? (2)



Quelle Illustration: Wikipedia



Beispiele Trainingsdaten für «KI» – mit willentlicher Handlung

- Personendaten aus Kundenstämmen, anderen Profildaten bei Diensten im Internet; über z. B. Ad-Firmen verknüpfbar
- Öffentliche Social-Media-Beiträgen in Text, Ton und (Bewegt-)Bild
- Standortprofilierung mit eingeschaltetem Handy
- Mobilitätsdaten bei Bikesharing (mit GPS-Tracking), bei modernen Autos mit SIM-Karten; gewissen ÖV-Angeboten
- Finanztransaktionen (bei Nutzung von Geldkarten, Apple Pay usw.)



Beispiele Trainingsdaten für «KI» – ohne willentliche Handlung (1)

- Werbettracking beim Surfen
- Videouberwachung im ÖV und öffentlichen Raum, inklusive Gesichtserkennung; auch Drohnen (Bereich: «IMINT»)
- Daten aus privater / staatlicher Massenüberwachung, wie z. B. E-Mails / Messaging-Direktnachrichten (Bereich: «COMINT» und «SIGINT»; vgl. Snowden-Leaks 2013)
- Datenabfluss durch Leaks bei Staat und Wirtschaft (nach Hacks)



Beispiele Trainingsdaten für «KI» – ohne willentliche Handlung (2)

TOP SECRET//SI//ORCON//NOFORN

Special Source Operations (TS//SI//NF) PRISM Collection Details

Current Providers

- Microsoft (Hotmail, etc.)
- Google
- Yahoo!
- Facebook
- PalTalk
- YouTube
- Skype
- AOL
- Apple

What Will You Receive in Collection (Surveillance and Stored Comms)?
It varies by provider. In general:

- E-mail
- Chat – video, voice
- Videos
- Photos
- Stored data
- VoIP
- File transfers
- Video Conferencing
- Notifications of target activity – logins, etc.
- Online Social Networking details
- Special Requests

Complete list and details on PRISM web page:
Go PRISMFAA

TOP SECRET//SI//ORCON//NOFORN

Quelle Illustration: NSA aus Snowden-Leaks (2013)



Manipulative Gefahren von «KI» DeepFakes speziell

- Systematischer Identitätsdiebstahl z. B. in Ausnutzung biometrischer (= nicht-geheimer) Personendaten
- Erfolgreichere Targeted-Angriffe mittels Laser-Phishing (Spear-Phishing maschinell auf Sprachstil und Inhalt «optimiert»)
- Systematische Erschaffung von – gefälschtem – Epressungsmaterial auf Basis privaten Datenmaterials, um z. B. herrschende Positionen zu stürzen
- Automatisierte Beeinflussung von Diskursen durch «All-Sicht» der Positionen; mittels realistisch sich verhaltender Akteure («Bots»)



Manipulative Gefahren von «KI» generell

- Gefahren individuell, wenn der «KI» zuviel Entscheidungsmacht im Alltag belassen wird; z. B. autonomes Autofahren bei unbekannten Mustern
- Totalüberwachung mit kafkaeskem Ausgang für Menschen basierend auf Gefahrenmustern / Social-Credits; à la «Black Mirror» (in naher Zukunft) bzw. der VR China (schon heute)
- Cyberangriffe durch automatisiertes Erlernen von systemischen Schwachstellen und Generierung entsprechender Angriffs-/Schadsoftware
- Automatische Waffensysteme ohne Moralkompass



«KI»-Regulierung in sinnvoll

- Rufe nach grundsätzlichem Technologieverbot oder auch Moratorien: unsinnig
- Regulierung / Transparenzschaffung von gewissen Anwendungen sicher sinnvoll: Algorithmen auf Social-Media, autonom fahrende Autos, Waffensysteme
- Verbote gewisser Anwendungen / Nutzungen sinnvoll, wo Privatsphäre / (Geschäfts-)Geheimnisse tangiert werden:
 - Beliebige Nutzung privater Daten, ohne jeden Consent
 - Gesichtserkennung im öffentlichen Raum à la VR China
 - Totalitäre Überwachung wie bei der EU-«Chatkontrolle»



Staat und Wirtschaft: generell dringender Handlungsbedarf!

- Abhängigkeiten Soft-/Hardware von Grossmächten auflösen – transparentere Systeme; weniger Kontrollverlust
- Automatisierte Ende-zu-Ende-Crypto auf dezentralisierter (P2P-)Infrastruktur; für real technische Datenverknappung
- Datenschutzverletzungen einklagbar und bussbar machen (ähnlich DSGVO / GDPR); Anreiz für sicherere Systeme
- Medienkompetenz in der Bevölkerung – besonders an Schulen – verbreiten
- Einsatz für Grund- und Freiheitsrechte – Digitalisierung nach Grundsätzen freiheitlich-demokratischer Ordnung!



Im Workshop: Kurze Anleitung zur Digitalen Selbstverteidigung



00_EDITORIAL	5
01_GRUNDLAGEN	10 DATENSPARSAMKEIT 11 MÄCHTIGE TECHKONZERNE 12 PASSWÖRTER 14 BETRIEBSSYSTEME 15 BACK-UP
02_DIE ALTERNATIVEN	18 WEBBROWSER 21 VPN 22 SUCHMASCHINEN 24 MESSENGER 26 SOZIALE NETZWERKE 29 E-MAILS 31 E-MAIL-PROGRAMME 32 KALENDER / ADRESSBUCH 33 ZUSAMMENARBEITEN IM NETZ 35 CLOUDDIENSTE UND ONLINESPEICHER 36 KARTEN 38 INTERNET SERVICE PROVIDER 39 HOSTING
03_GLOSSAR	41

Quelle Ratgeber: <https://digiges.ch/ratgeber>

