Diskussionen zu KI-Systemen, ethische Grenzen und Auswirkungen

17.11.2016

Meister Rados

Übersicht

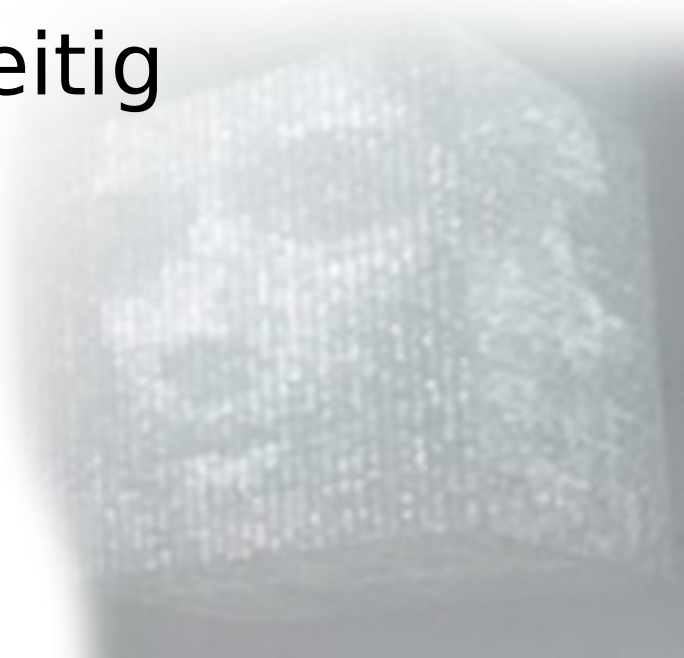
- Einleitung: Problemstellung – Ein Spiel
- Kontrolle
- AI Guardians
- Turing‘s Red Flag

Ein Spiel

- Deep Mind‘s AlphaGo vs. Lee Sedol endet 4:1
- AlphaGo spiel Go besser als seine Programmierer
- Wahrscheinlich sogar besser als jeder Mensch
- Scheinbar niemand versteht, was AlphaGo versteht

Kontrolle

- AI Systeme werden zunehmend unsere Alltagsgegenstände
- Sie treffen für uns Entscheidungen und übernehmen Verantwortung
- Die Fehlerfreiheit dieser Systeme wird immer wichtiger und gleichzeitig schwieriger zu verifizieren

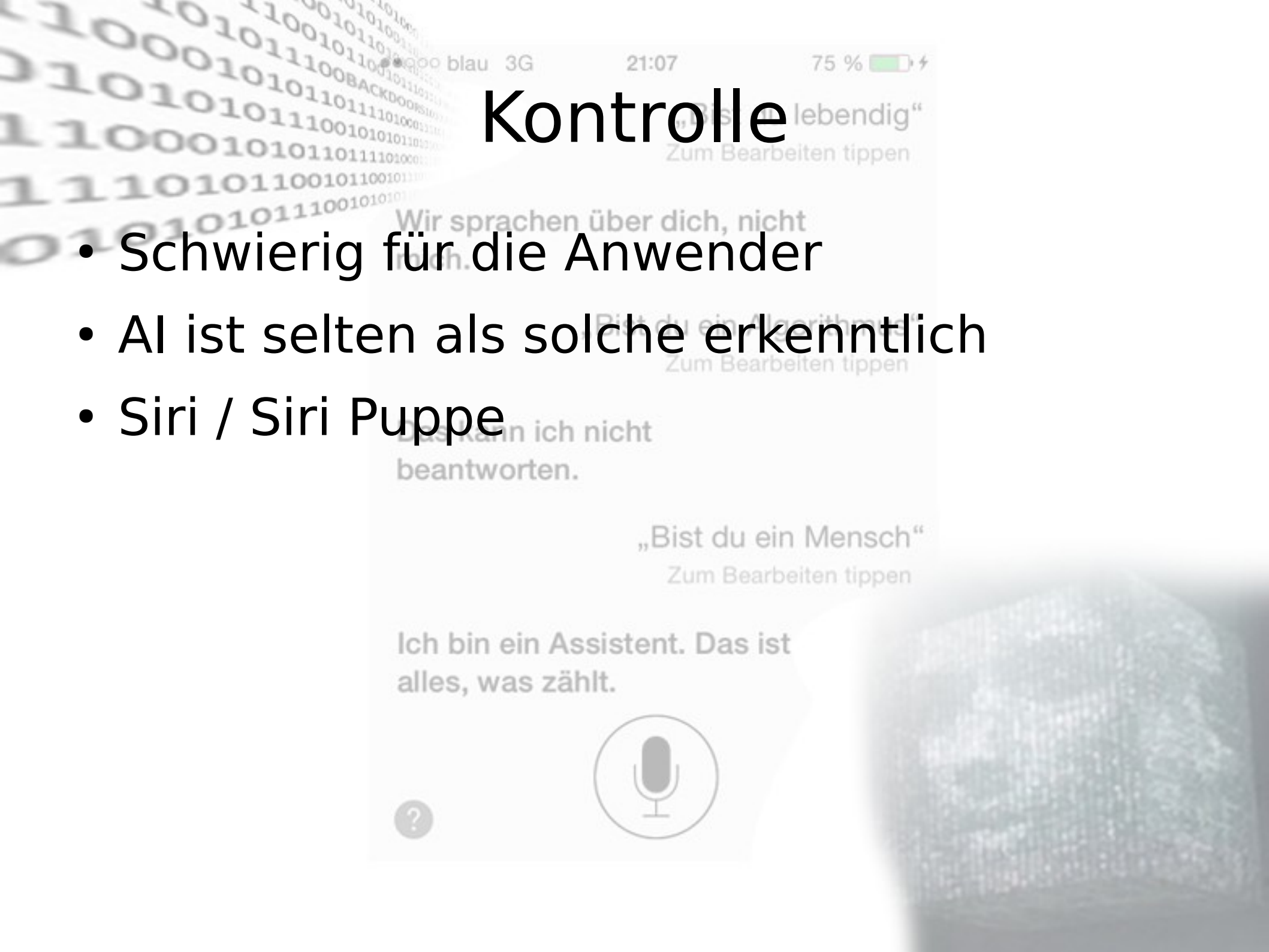


Kontrolle

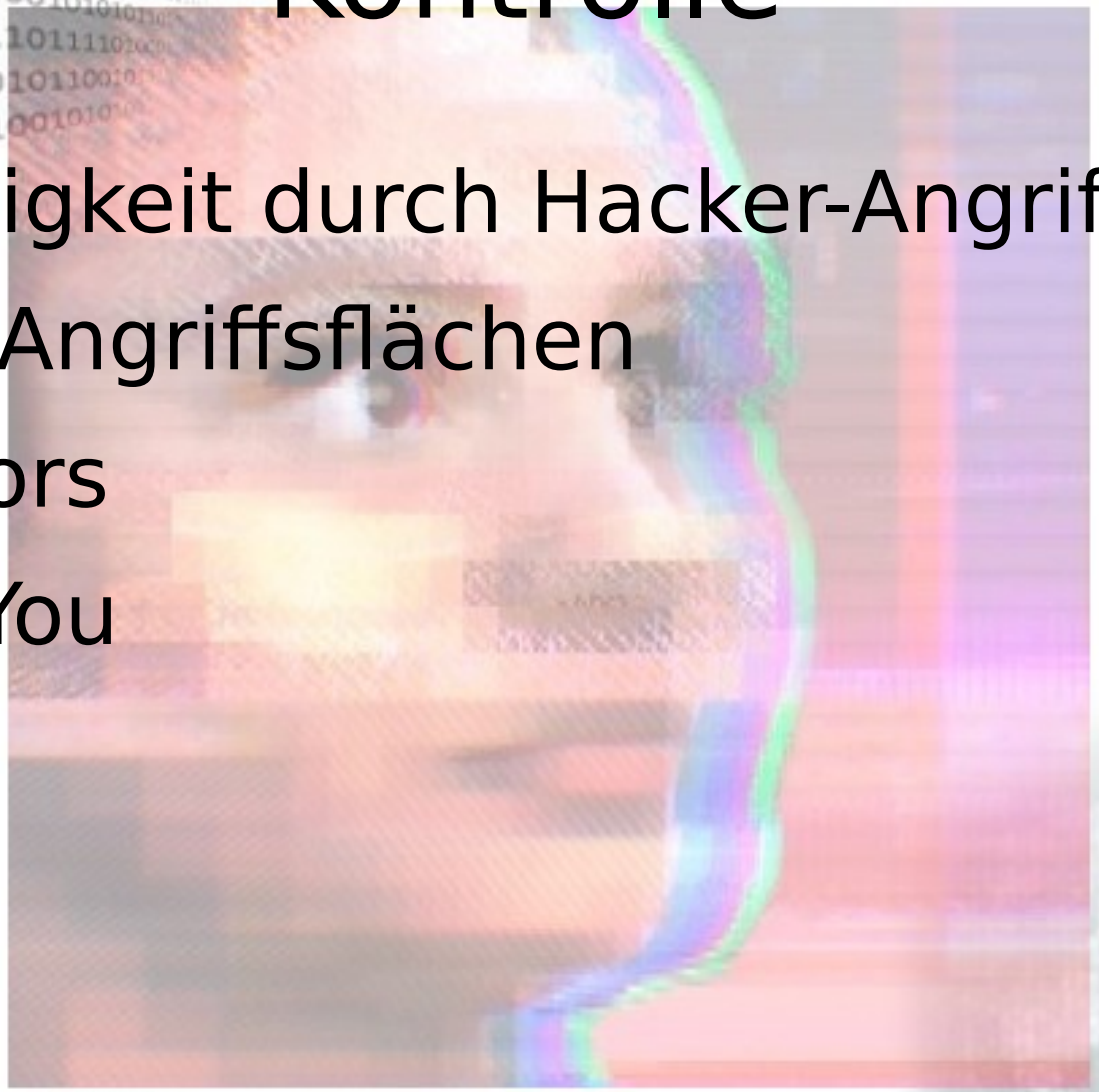
- Schwierig für die Programmierer
- Obfuskation
- Zu komplex oder zu kompliziert
- Zu viele verschiedene Sprachen / Libraries / Stile
- → Black Box, die Entscheidungen fällt
- Gedankenspiel: Anti-Malware AI vs. Malware AI

Kontrolle

- Schwierig für die Anwender
- AI ist selten als solche erkenntlich
- Siri / Siri Puppe



Kontrolle



- Schwierigkeit durch Hacker-Angriffe
- Andere Angriffsflächen
- Backdoors
- TayAndYou

AI Guardians

Features

- Idee / Entwurf dass AI von AI kontrolliert und geleitet wird
- Interagieren nur indirekt mit unserem Umfeld
- Halten agierende AI Systeme gegebenenfalls „auf Kurs“
- Kennen die Intention des Programmierers und unsere natürlichen Rechte und Normen sowie ethische Werte

AI Guardians

Features

- Können „on the fly“ angepasst werden
- Sollten mit dem Lerntempo der agierenden AI Systeme mithalten können und Fehlerquellen der agierenden AI Systeme lokalisieren

AI Guardians

Bedingungen an agierende AI

- Agierende AI Systeme müssen angepasst werden
- Sie müssen auf ihre Gaurdians hören und dürfen diese nicht aushebeln
- Sie brauchen einen zuverlässigen on/off switch und zuverlässige fail-safe Routinen

AI Guardians

Umsetzung

- Wer schreibt den Quellcode?
- Wer legt die Werte und Normen fest, an denen sich AI Guardians orientieren sollen?
- Wie formt man daraus für Maschinen verständlichen Input?
- Inwiefern sollte der Nutzer die Gaurdians anpassen können?
- Die Rechtslage müsste zugeschnitten werden
- Internationaler Konsens

Turing‘s Red Flag

- Idee AI Systeme als solche erkenntlich zu machen
- Onlinespiele – Bots (Poker, Shooter, …)
- Textgenerierung
- Siri Puppe
- Autonome Fahrzeuge

Turing‘s Red Flag
Umsetzung

- Klares Statement vor Beginn der Interaktion
- Die Interaktion darf in ihrer Form nicht einer Menschlichen gleichen
- Sehr Kontextbezogen (z.B. Andere Kennzeichen, ...)
- Ist es zu spät / zu früh?

Ausblick

Fragen?

Danke für eure
Aufmerksamkeit!