

Der Turing Test und seine Varianten

Woran erkennt man intelligentes Verhalten?

Julian Bucher

Abstrakt

Menschen und Maschinen zu unterscheiden ist eine Aufgabe vor welcher manche Unternehmen täglich stehen. Ein bekanntes Mittel hierfür sind zum Beispiel Captchas welche gerade dafür entwickelt wurden um Maschinen bzw. Bots oder Spam zu erkennen. Man trifft auf Captchas zum Beispiel wenn man sich einen neuen E-Mail Account anlegen möchte, denn ein Unternehmen wird kaum wollen, dass ein Bot sich auf die schnelle ganz viele neue E-Mail Accounts anlegt.

Man kann allerdings auch die Frage anders herum stellen und fragen, wann man Menschen und Maschinen nicht mehr unterscheiden kann. Wie kann man eine Maschine erkennen, die denken kann wie ein Mensch - eine Maschine die intelligent ist? Woran erkennt man also intelligentes Verhalten bei einer Maschine und wie lange kann man die Intelligenz eines Menschen von der einer Maschine unterscheiden?

Hierzu betrachten wir das Imitation Game von A. M. Turing und die Anfänge von Chatbots. Auch werden Kritikpunkte am Imitation Game untersucht und zwei sich daraus entwickelte Alternativen zum Imitation Game, nämlich die Winograd Schema Challenge und der Lovelace Test 2.0 näher beleuchtet.

Das Imitation Game

Im Jahr 1950 beschreibt A. M. Turing sein Imitation Game [Tur], welches sich wie folgt darstellen lässt.

Drei Personen nehmen am Imitation Game teil. Ein Mann, eine Frau und eine weitere Person, egal welchen Geschlechts, im folgenden Jury genannt. Die Jury ist dabei räumlich getrennt von den beiden anderen Personen.

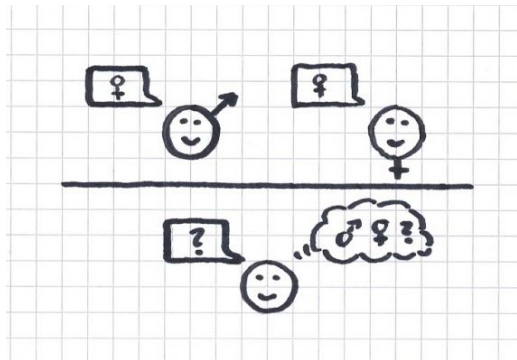
Das Ziel der Jury ist herauszufinden, welcher der beiden Anderen der Mann und wer die Frau ist. Dafür sind Mann und Frau für die Jury wahllos mit einer Markierung, zum Beispiel X und Y versehen, so dass nach einer vorher festgelegten Zeitspanne, die Jury am Ende des Spiels entweder sagt X ist der Mann und Y die Frau, oder X ist die Frau und Y der Mann.

Um dieses Ziel zu erreichen darf die Jury mit X und Y, in einer Form welche keinerlei Rückschlüsse auf das Geschlecht von X bzw. Y zulässt, kommunizieren und so zum

Beispiel Fragen stellen. Turing schlägt hier eine getippte schriftliche Kommunikation vor.

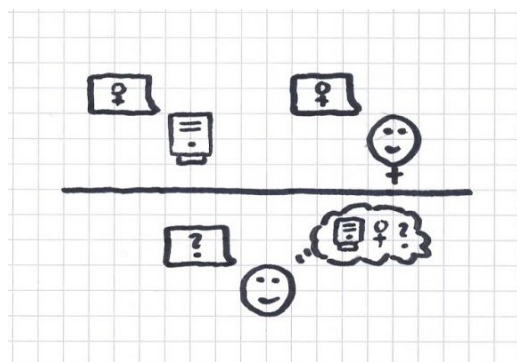
Das Ziel des Mannes ist es, dass sich die Jury nicht entscheiden kann, wer X und wer Y ist. Er wird also nicht immer die Wahrheit sagen und versuchen, dass die Jury denkt, er sei die Frau. Das Ziel der Frau ist es, dass sich die Jury am Ende des Imitation Games richtig entscheidet.

Abb. 1 - Imitation Game Version 1



Im zweiten Schritt ersetzt Turing den Mann durch einen Computer und stellt die Frage, ob die Jury sich noch genauso oft richtig entscheidet, wer X und wer Y ist, wie es galt sich zwischen Mann und Frau zu entscheiden.

Abb. 2 - Imitation Game Version 2



Das Ziel von Turing mit seinem Imitation Game war die Frage zu beantworten, wann Computer denken. Diese Frage hat laut Turing die Schwachstelle, dass man erst einmal eine genaue Definition des Begriffes "Denken" vornehmen muss, was nicht einfach ist. Turing ist der Meinung, dass sein Imitation Game die Frage adäquat ersetzt, da die Fragen, welche die Jury stellen kann, aus so vielen unterschiedlichen Bereichen kommen können und zur Beantwortung dieser somit eine sehr große Spannbreite menschlicher Denkprozesse zum Einsatz kommen muss.

Ein Beispiel von Turing für einen möglichen Ausschnitt aus einer solchen Konversation wäre das Folgende:

Abb. 3 - Mögliche Konversationen im Imitation Game

Frage: Kannst Du mir ein Sonett über die Forth Brücke schreiben?
Antwort: Lass mich da raus. Ich konnte niemals Gedichte schreiben.
Frage: Okay. Spielst Du Schach?
Antwort: Ja!
Frage: Ich habe K auf H1 und sonst nichts mehr. Du K auf H4 und L auf E3. Was ist dein Zug?
Antwort: K auf H3, Kumpel.

Chatbots

1966 entwickelte Joseph Weizenbaum ein Programm namens ELIZA, welches es Menschen ermöglichte mit einer Maschine in schriftlicher Form zu kommunizieren, also einen Chatbot. ELIZA kann verschiedene Gesprächspartner simulieren, unter anderem auch einen Psychologen.

Die Funktionsweise von ELIZA ist dabei relativ simpel und wird wie folgt von Saygin [Say] wiedergegeben: Der eingegebene Text wird von ELIZA durchsucht, ob dieser ein für ELIZA bekanntes Wort enthält. Ist dies der Fall wird ein vorgefertigter Satz zurückgegeben. Enthält der eingegebene Text kein für ELIZA bekanntes Wort, so wird eine ausweichende Antwort wie, "Tut mir leid, das verstehe ich nicht. Lass uns über etwas anderes reden." gegeben.

Weizenbaum selbst begründet mit der Funktionsweise ELIZAs gerade die Entscheidung, dass ELIZA einen Psychologen darstellt, da ein Psychologe selbst nicht viel über sich erzählt und reflektierende Fragen über das was ihm erzählt wurde stellt. Genau das, was ELIZAs Aufbau entgegenkommt. ELIZA reflektiert auch gerade so die Aussagen, welcher man ihr erzählt.

Sagt man ELIZA zum Beispiel: "Ich verstehe mich nicht gut mit meiner Schwester." wäre eine mögliche Antwort ELIZAs: "Erzähle mir mehr über deine Schwester." Personen die mit ELIZA kommunizierten erwarteten auch mit einem Psychologen zu sprechen und waren deswegen auch nicht erstaunt über ELIZAs einseitige Antworten.

Dies ging so weit, dass manche Personen, welche mit ELIZA kommunizierten davon ausgingen mit einem wirklichen Psychologen zu sprechen und nicht mit einem Chatbot.

ELIZA hätte niemals einen richtiges Imitation Game bestehen können. Sie war allerdings ein Beleg dafür, dass Chatbots in der Lage sind spezielle Arten von Gesprächen und Aspekte menschlicher Kommunikation zu imitieren.

Ein ähnlicher Ansatz war das Chatprogramm PARRY aus dem Jahr 1972 von Kenneth Colby. PARRY verfolgte den Ansatz dass er seine Gegenüber nie zu Wort kommen lässt und stets seine vorgefertigte Geschichte erzählen will. Zu diesem Zweck stellt PARRY einen paranoiden Schizophrenen dar.

Wie Saygin beschreibt wurde eine abgewandelte Variante des Imitation Games durchgeführt, in welchem erfahrene Psychologen versuchten PARRY von einem wirklichen Paranoiden zu unterscheiden. Laut Saygin hatten viele Psychologen Schwierigkeiten zu entscheiden, wer PARRY war und wer ein wirklicher Mensch. Eine Folge dieser ersten Chatbots war der Loebner Preis [Loe] welcher seit 1991 existiert, welcher sich selbst als ältester Turing Test Wettbewerb bezeichnet. Beim Loebner Preis hat man ein Zeitlimit von 25 Minuten in welchem 4 Jurymitglieder jeweils mit einem Chatbot und einem Menschen kommunizieren. Nach Ablauf des Zeitlimits entscheidet sich jedes Jurymitglied welcher der beiden Chatbot und wer Mensch war. Der erste Gewinner des Loebner Preises war ELIZA und der aktuelle Gewinner [aus dem Jahr 2016] der Chatbot Mitsuku. Alle bisherigen Gewinner des Loebner Preises würden den Turing Test nicht bestehen. Sie haben den Preis gewonnen, denn sie waren der beste aktuelle Chatbot, welcher "am nächsten dran" ist den Turing Test zu bestehen, auch wenn dies noch "sehr weit weg" bedeuten kann.

Kritik am Imitation Game

Das Imitation Game welches Turing 1950 vorstellte wurde häufig aus unterschiedlichsten Gründen kritisiert.

Ein Argument wäre, dass Computer vielleicht etwas tun, was Denken sehr nahe kommt, sich aber immer noch grundsätzlich davon unterscheidet, wie eine Person denkt. Vielleicht haben Computer einfach eine andere Struktur um zu denken, welche nicht mit den menschlichen Denkstrukturen übereinstimmt.

Saygin geht hier auf den Seagull Test (Möwen Test) ein. Vereinfacht möchte man bei diesem Test untersuchen, ob etwas die Fähigkeit besitzt zu fliegen. Das einzige bekannte Objekt bei Erstellung des Testes, welches die Fähigkeit des Fliegens besitzt, ist die Möwe. Daraufhin stellt man Beobachtungen an, um herauszufinden, wie die Möwe fliegt. Sicherlich wird man zu Ergebnissen wie, die Möwe schlägt mit den Flügeln, oder Ähnlichem kommen.

Würde man also Flugzeuge dem Seagull Test unterziehen, würde man zum Ergebnis kommen, dass Flugzeuge nicht fliegen können, denn sie schlagen nicht mit den Flügeln.

Zusammengefasst könnte man sagen, dass ein Computer zu sehr eine Person imitieren muss, um das Imitation Game zu bestehen und er nicht die Möglichkeit hat eine eigene Möglichkeit des Denkens zu entwickeln.

Als weiteren Kritikpunkt beim Imitation Game, kann man die Frage aufwerfen, ob eine Konversation überhaupt die beste Art von Test ist, vor allem mit menschlichen Jurys, da sich diese auch irren oder getäuscht werden können, wie es zum Beispiel bei ELIZA oder PARRY geschah.

Aus diesen Gründen entwickelte Hector Levesque eine alternative Form des Turing Tests, die Winograd Schema Challenge [Lev].

Winograd Schema Challenge

Die Winograd Schema Challenge verfolgt die folgenden Grundgedanken:

- Das Testobjekt soll auf eine große Auswahl verschiedener englischer Sätze antworten.
- Erwachsene englische Muttersprachler sollen in der Lage sein den Test einfach zu bestehen.
- Wenn der Test bestanden ist soll gesagt werden können, dass das Testobjekt denken kann.

Bei der Winograd Schema Challenge muss das Testobjekt auf Fragen mit zwei Antwortmöglichkeiten antworten und dabei die richtige Antwort wählen. Die vorkommende Worte in den Fragen und Antwortmöglichkeiten stammen dabei aus einem "Wortpool" mit verschiedenen englischen Wörtern. Dieser Wortpool ist dem Testobjekt bereits vor dem Test bekannt. Des Weiteren muss jede Frage die folgenden vier Eigenschaften erfüllen:

1. Zwei verschiedene Parteien kommen jeweils in einer Nominalphrase vor. Die Parteien können sowohl Objekte, oder auch Personen, etc. sein.
2. Ein Pronomen oder Possessivadjektiv bezieht sich auf eine der beiden Parteien. Grammatikalisch hat es die Form, so dass es sich auf beide Parteien beziehen könnte.
3. Die Beantwortung der Frage erfordert die Zuordnung von 2., also dem Possessivadjektiv bzw. Pronomen, zur richtigen Partei. Die erste Antwort auf die Frage ist immer die zuerst erwähnte Partei, die zweite Antwort die als zweites erwähnte Partei.
4. Es gibt ein spezielles Wort (special word genannt) welches, durch ein spezielles anderes Wort (alternate word genannt) ersetzt, dazu führt, dass die andere Antwortmöglichkeit die richtige Antwort ist.

Die Wahl der richtigen Antwort erfordert daher das genaue verstehen des Kontextes in welchem die Frage gestellt ist und der vierte Punkt schließt ein einfaches analysieren des Fragenschemas aus. Nach Levesque ist es daher notwendig "denken zu können", wenn man die richtige Antwort geben will.

Aufgrund der grammatikalischen Eigenschaften der deutschen Sprache ist es im Deutschen meist so, dass sich wegen des Genus die zweite Eigenschaft für Fragen im Deutschen nicht (einfach) erfüllen lässt.¹

Eine Fragenbeispiel von Levesque ist das Folgende:

Abb. 4 - Fragenbeispiel Winograd Schema Challenge

The trophy would not fit in the brown suitcase because it was too big.

What was too big?

0: the trophy

1: the suitcase

Die beiden Parteien sind hierbei "the trophy" und "the suitcase". Das Wort "it" erfüllt die zweite Bedingung und "big" ist das special word mit dem alternate word "small". In diesem Beispiel ist es notwendig zu wissen, dass man in einen Koffer etwas hineinfüllen kann und wenn etwas nicht in den Koffer passt, weil etwas zu groß ist, dass dann nicht der Koffer zu groß ist, sondern das, was in den Koffer passen soll. Die größte Schwierigkeit beim Erstellen einer Winograd Schema Challenge ist es sicherlich, geeignete Fragen zu finden. Fragen sollten natürlich stets die oben genannten vier Bedingungen erfüllen, aber es sollte auch nicht zu offensichtlich sein, was die richtige Antwortmöglichkeit ist. Ist in der Frage zum Beispiel eine der Parteien ein Rennauto und es wird danach gefragt, welche Partei zu schnell war, so liegt die Antwort Rennauto nahe, auch wenn man die vollständige Frage nicht kennt. Auf der anderen Seite ist es aber auch notwendig, dass es aus logischer Sicht nur eine korrekte Antwortmöglichkeit gibt. Wenn eine Antwortmöglichkeit nur sehr unwahrscheinlich ist, heißt das nicht, dass es nicht Fälle geben könnte, in welchem sie zutrifft. Im folgenden Beispiel ist Antwort 1 die offensichtliche, aber aus rein logischer Sicht, könnte auch Antwortmöglichkeit 0 zutreffen.

Abb. 5 - Fragenbeispiel Winograd Schema Challenge

The town councillors refused to give the angry demonstrators a permit because they advocated violence. Who advocated violence?

0: the town councillors

1: the angry demonstrators

¹ Im englischen Satz "The trophy would not fit in the brown suitcase because it was too big." kann sich das Wort "it" aus grammatikalischer Sicht sowohl auf "trophy" als auch auf "brown suitcase" beziehen. Die letztendliche Bedeutung wird erst aus dem Kontext klar.

Übersetzt man hingegen den Satz ins Deutsche: "Die Trophäe würde nicht in den braunen Koffer passen, denn sie war zu groß.", so wird das "it" zu einem "sie". Allerdings ist nur "die Trophäe" eine weibliche Partei und "der braune Koffer" männlich. Es ist also ohne den Kontext zu kennen allein von der grammatikalischen Struktur her klar, dass sich das "sie" auf "die Trophäe" bezieht und nicht auf "den braunen Koffer".

Lovelace Test 2.0

Teilweise wird argumentiert, dass ein Computer nur wirklich denken könne, wenn er in der Lage ist, selbst etwas zu erschaffen, oder andere zu überraschen. Diese Begründung ist durchaus legitim, wenn man bedenkt, dass es durchaus möglich ist ein gewisses Maß an Kreativität als Voraussetzung für Intelligenz zu sehen.

Hier setzt der Lovelace Test und der Lovelace Test 2.0 von Mark Riedl [Rie] an, welcher eine Erweiterung des original Lovelace Tests aus dem Jahr 2001 von Bringsjord et al darstellt. Hierbei wurde es sich zum Ziel gesetzt, "erschaffen" und "überraschen" genauer zu definieren.

Der Lovelace Test 2.0 wird von Riedl wie folgt beschrieben:

- Ein Testobjekt A muss ein Artefakt O vom Typ T erschaffen.
- Dabei muss O einer Menge von Bedingungen C gehorchen, wobei die Elemente aus C jede mögliche von Menschen formulierte Bedingung sein können.
- Ist eine Person H, welche C und T gewählt hat² davon überzeugt, dass O zu C und T passt, so gilt der Lovelace Test 2.0 als bestanden.

Ein Beispiel für eine Aufgabe von H könnte die folgende Aufgabe sein:

"Komponiere ein Klavierstück in A-Dur, welches an einen kalten Wintertag erinnert."

Levesque vertraute gerade nicht menschlichen Jurys, da diese wie im Fall von ELIZA getäuscht werden könnten und entwickelte so einen von Jurys unabhängigen Test.

Riedl hingegen legt beim Lovelace Test 2.0 gerade Wert darauf, dass hier menschliche Jurys Vorteile haben und nicht getäuscht oder überlistet werden können. Dies ist seiner Meinung nach der Fall, da gerade die Person H das C und T wählt, so dass A den Lovelace Test 2.0 aus seiner Sicht gerade nicht bestehen kann. H ist dann natürlich überrascht, sollte A diesen doch bestehen. Auch kann H selbst frei entscheiden, wann A seine gestellten Bedingungen C erfüllt hat und ob der Test bestanden ist. Sollte A den Test bestehen hätte H so später die Möglichkeit C solange zu verschärfen, bis A den Test nicht mehr besteht. Hieraus könnte man dann ableiten, wie überrascht H von A ist.

Captchas

Das Wort Captcha [Deutsch: Fangen] steht für **C**ompletely **A**utomated **P**ublic **T**uring **T**est to tell **H**umans and **C**omputers **a**part. Wie zu Beginn erwähnt, werden Captchas allerdings gerade dazu eingesetzt um zwischen Mensch und Maschine zu unterscheiden. Es wird also das gegenteilige Ziel des Turing Tests verfolgt. Man möchte sagen können, wer die Maschine ist und wer der Mensch und nicht wie beim Turing Test es schaffen, dass man gerade diese Unterscheidung nicht mehr vornehmen kann.

² Von einer unabhängigen dritten Person R wird zuvor überprüft, ob bei gewählten C und t die Aufgabe für einen normalen Menschen erfüllbar ist.

Über die Jahre haben sich Captchas stetig weiterentwickelt. Die ersten Captchas waren reiner verschnörkelter Text, welcher gelesen und richtig eingegeben werden musste. Die Firma Google reCaptcha [Goo], verwendet als Testmaterial zum Beispiel Wörter aus der New York Times, welche Computer beim einscannen des Archives der New York Times nicht einwandfrei erkennen konnten.

Mit der Zeit wurde die Texterkennung allerdings immer besser und besser und Google selbst behauptete bereits im Jahr 2014, dass sie es schaffen 99,8% der Captchas, welche auf verschnörkelte Schrift basieren, knacken zu können.

Die Schrift wurde also immer verschnörkelter, bis auch für den Menschen extrem schwer wurde den eigentlichen Text zu entziffern.

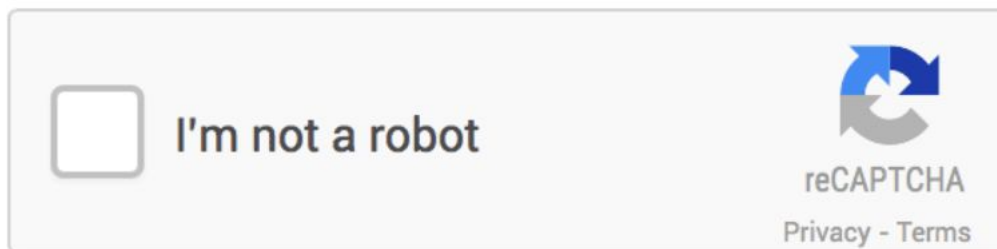
Die nächste Art von Captchas welche sich als Nachfolger der verschnörkelten Schrift etablierte beruhte auf Bilderkennung. Es werden verschiedene Fotos gezeigt und man muss beispielsweise alle Fotos auswählen, welche einen Kuchen zeigen.

Alternativ gibt es auch Captchas, welche ein Foto einer Zahl zeigen und man muss eingeben, welche Zahl auf dem Foto zu sehen ist.

Dadurch, dass allerdings auch die Bilderkennung immer weitere Fortschritte macht, werden auch immer mehr solcher Captchas von Bots geknackt.

Google reCaptcha hat daraufhin die I am not a robot Captcha entwickelt, die derzeit (2017) zu den neuesten Arten von Captchas gehört.

Abb. 6 - No Captcha Recaptcha Captcha [Bild von Goo]



Hierbei reicht ein einfacher Klick in die Box und Google entscheidet, ob es sich um einen Menschen oder eine Maschine handelt. Nur wenn Google sich nicht sicher ist, kommt zusätzlich eine auf Bilderkennung basierende Captcha zum Einsatz.

Google entscheidet hierbei auf Grundlage verschiedener Daten. Zuerst wird die Mausbewegung innerhalb der Box verfolgt und welchen Weg der Mensch bzw. die Maschine bis zum Anklicken mit der Maus zurücklegt. Weiterhin greift Google auf verschiedene andere Informationen zurück, wie etwa die IP-Adresse oder die verwendeten Cookies des Benutzers, um zu entscheiden, ob es sich nun um einen Menschen oder eine Maschine handelt. Details, wie genau dies funktioniert, teilt Google der Öffentlichkeit allerdings nicht mit.

Man könnte sich nun die Frage stellen, ob Captchas nicht ausreichen um zu entscheiden, ob es sich um eine intelligente Maschine handelt. Die Antwort hierauf ist allerdings sicherlich ein "Nein". Nur weil ein Bot in der Lage ist eine Captcha zu knacken, was in der Vergangenheit häufig schon vorkam, heißt es nicht, dass dieser intelligent ist.

Allgemein bezeichnet man Probleme im Bereich der künstlichen Intelligenz als AI-Complete oder auch AI-Hard, wenn die Lösung des Problems äquivalent zum Problem des Erschaffens eines Computers mit menschlicher Intelligenz ist.

Damit sind unsere jetzigen Captchas sicherlich nicht AI-Complete. Das Ziel eine Maschine zu bauen, welche den Turing Test besteht wäre beispielsweise ein AI-Completes Problem.

Abschluss und eigene Meinung

Möchte man wirklich die Fragestellung klären, woran man intelligentes Verhalten erkennt, ist der Turing Test sicherlich ein guter Ansatz. Jedoch ist gerade der Kritikpunkt, dass eine Maschine, die den Turing Test bestehen will, zu sehr ein Mensch sein muss bzw. diesen imitieren muss, für mich ein Hauptargument gegen den Turing Test. Nur weil Computer extrem schnell rechnen können, sollten sie dies zum Beispiel bei einem Turing Test verheimlichen?

Sicherlich deckt der Turing Test allerdings auch unter den hier vorgestellten Verfahren die größte Spannweite an unterschiedlichen Gebieten im Zusammenhang mit künstlicher Intelligenz ab.

Der Lovelace Test verfolgt einen interessanten Ansatz und mit diesem kann man meiner Meinung nach auch gut Kreativität messen, was ihn für mich auch zu einem sinnvollen Test macht.

Dennoch denke ich, dass dieser heutzutage fast schon überflüssig ist, denn es ist aus meiner Sicht gelungen, Computern Kreativität beizubringen.³

Möchte man letztendlich klären, was genau man unter Intelligenz bzw. Denken versteht, sind der Turing Test aber auch die Winograd Schema Challenge sicherlich ein guter Ansatz, aber man wird nicht wirklich weiter kommen, wenn man sich davor nicht darauf einigt, was genau es bedeutet intelligent zu sein.

Hierbei kommt auch immer ein interessantes Paradoxon vor:

Lange Zeit galt man als intelligent, wenn man gut Schach spielen konnte. Als dann ein Schachcomputer erstmals den Weltmeister besiegte, kam schnell die Meinung auf, der Schachcomputer sei nicht intelligent, denn er habe einfach "Brut Force" verschiedene Möglichkeiten ausprobiert und die Beste ausgewählt. Das gleiche

³ In A Neural Algorithm of Artistic Style von Leon A. Gatys, Alexander S. Ecker und Matthias Bethge wird beispielsweise beschrieben, wie es gelungen ist verschiedene Kunststile einem neuronalen Netz beizubringen, welches diese dann auf beliebige Bilder übertragen kann. Als weiteres Beispiel hat sich Google DeepMinds selbst das Laufen beigebracht und kam aus meiner Sicht zu sehr erstaunlichen Ergebnissen.

passierte beim Go, ähnliches mit dem Computer welcher das Spiel Jeopardy gewann.

Gilt also etwas nur so lange als intelligent, wie es kein Computer bzw. keine Maschine gibt, die in der Lage ist, dies auch zu können? Gilt somit eine Maschine quasi nie als intelligent?

Ich denke die Menschen müssen sich damit abfinden, dass es irgendwann Computer/Maschinen geben wird die intelligent sind, was auch immer man sich genauer darunter vorstellt, wenn man den Begriff "intelligent" nicht alleine für den Menschen reserviert.

Turing selbst sagte 1950 das Folgende für das Jahr 2000 voraus:

"I believe that in about fifty years' time it will be possible, to programme computers (...) to make them play the imitation game so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning. The original question, "Can machines think?" I believe to be too meaningless to deserve discussion. (...)"

Sicherlich ist dies so noch nicht eingetreten und man sollte natürlich auch nicht neue Jahreszahlen nennen, wann dies der Fall sein wird. Ich bin allerdings der festen Überzeugung, dass Turings Vorhersage irgendwann zutreffen wird und der Turing Test, trotz all seiner Schwächen, auf diesem Weg - als Ziel für künstliche Intelligenz - eine wichtige Rolle spielen wird, wenn man entscheiden will, wann es soweit ist.

References

[Goo] Vinay Shet, "Are you a robot? Introducing "No CAPTCHA reCAPTCHA"" am 3. Dezember im Jahr 2014 im Google Webmaster Central Blog

[Lev] Hector J. Levesque, "The Winograd Schema Challenge" im Jahr 2011 American Association for Artificial Intelligence

[Loe] Hugh Loebner, <http://loebner.net/Prize/loebner-prize.html>

[Rie] Mark O. Riedl, "The Lovelace 2.0 Test of Artificial Creativity and Intelligence" im Jahr 2014

[Say] Saygin et al, "Turing Test: 50 Years Later" im Jahr 2000 in Minds and Machines 10 Seiten 463-518

[Tur] Turing, A. M., "Computing machinery and intelligence" im Jahr 1950 in Mind 59 Seiten 433-460