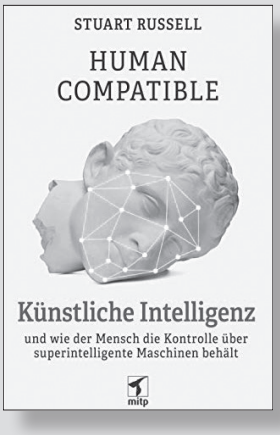
Dorothea Kolossa

Stuart Russell: Human Compatible – Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält

Buchbesprechung

Die aktuellen Entwicklungen in der künstlichen Intelligenz werden von vielen (auch von mir) als existentielles Risiko für die Menschheit und das Leben auf unserem Planeten gesehen. Gleichzeitig besitzt Künstliche Intelligenz (KI) ein enormes Potenzial zum Besseren. Mit guten Modellen kann sie nicht nur große Arbeitserleichterungen bringen (denn wer will wirklich selbst mit Excel arbeiten oder stundenlang einen Laster über die Autobahn steuern?) sondern auch in Wissenschaft und Gesellschaft bei der Analyse komplexer Fragen helfen, und uns beim Treffen intelligenterer Entscheidungen auf solider Basis von Daten und Erkenntnissen unterstützen. Damit kann sie auch helfen, andere existentielle Risiken einzudämmen und viele wichtige Entwicklungen – in der Medizin und Biologie, der Umwelttechnik, der Physik und den Ingenieurwissenschaften – wesentlich schneller voranzubringen, als es ohne diese Werkzeuge möglich wäre.

In seinem aktuellen Buch *Human Compatible* beschreibt Stuart Russell die existentiellen Risiken einer künstlichen Intelligenz, die mit aktuellen Methoden und Kostenfunktionen entworfen wird. Er beschreibt aber auch eine Strategie, um künstlich intelligente Systeme so zu gestalten, dass sie nicht dazu prädestiniert sind, die Zukunft der Menschheit zu vernichten. Damit rücken die Möglichkeiten einer positiven künstlichen Intelligenz und eines an menschlichen Bedürfnissen ausgerichteten maschinellen Lernens in den Vordergrund.



Stuart Russell
Human Compatible
Künstliche Intelligenz und wie der Mensch die Kontrolle über superintelligente Maschinen behält
mitp Verlags GmbH, 2020
376 Seiten
Preis € 15,00 (Taschenbuch)
ISBN: 978-3-7475-0173-3

Erstes Thema: Warum KI ein existentielles Risiko darstellt

Die Fortschritte in der Entwicklung *intelligenter* Systeme sind beeindruckend und recht rasant. KI-Systeme können Sprache und Bilder bei entsprechendem Training ähnlich gut erkennen wie Menschen und GPT-3 schreibt inzwischen schönere Texte als manch einer von uns – jedenfalls an unseren weniger guten Tagen. Maschinelles Lernen und künstlich intelligente Systeme optimieren aber in ihrem Verhalten immer formal gefasste sogenannte *Kostenfunktionen*. Beispielsweise könnte ein KI-System zur Routenplanung immer versuchen, die Reisedauer zu minimieren – die Aufgabenstellung wäre dann einfach:

> *Verhalte Dich so, dass die Kostenfunktion F = Reisedauer minimal wird.*

Leider ist es aber sehr schwer, eine formale Kostenfunktion aufzustellen, die eine (zukünftige) superintelligente KI sinnvoll optimieren kann, ohne die Menschheit zu vernichten, oder jedenfalls mindestens krassen Schaden anzurichten. Das einfachste von vielen Beispielen einer derartigen Fehloptimierung in Human Compatible ist König Midas. Er wünschte sich, dass alles, was er berührt, zu Gold wird. Leider hatte er nicht bedacht, dass man Gold bekanntermaßen nicht essen kann. Ein noch fataleres Beispiel, mit konkreterem Bezug zu intelligenten Maschinen: Die Lösung der Optimierungsaufgabe, menschliches Leiden zu minimieren, führt in einen stabilen Punkt, an dem das menschliche Leiden genau Null ist.

Aber auch, wenn man versucht, die Kostenfunktionen einer KI mit allergrößter Sorgfalt zu gestalten, läuft man immer wieder in dieselbe Falle: Immer wieder fehlen wichtige Aspekte in der Kostenfunktion, so dass letztlich die engstirnige Optimierung derselben direkt in eine Katastrophe führt.

Aufgabe zur Illustration: Überlegen Sie sich eine formale Kostenfunktion, die in letzter Konsequenz und mit vollem Einsatz optimiert werden kann, ohne dass existentielle Probleme entstehen. Leider wird man eine superintelligente KI auch nicht gut abschalten können, wenn sie einmal damit loslegt, ihre Kostenfunktion zu optimieren – denn wenn sie ausgeschaltet ist, wird die Kostenfunktion ja nicht mehr optimiert. Also muss sie die Menschen auch von ihrem Ausschalter fernhalten, um die Kosten zu minimieren! Und wie aus Spielen wie AlphaGo schon aktuell für einfache KI und kleine Kunstwelten klar ist, kann eine KI wesentlich weiter in die Zukunft planen, als irgendein Mensch, der sie gerne ausschalten möchte. Bisher ist das vergleichsweise unproblematisch, weil auch die intelligentesten Systeme noch weitreichenden technischen (oder kognitiven?) Beschränkungen unterliegen. Aber einerseits ändern sich viele dieser Einschränkungen heute schnell – sogar diejenigen, die eine Bewaffnung von KI-Systemen verbieten sollten – und andererseits kommen auch Systeme aus Maschine und menschlichem *Handler* mit ähnlichen Problemen daher, da in vielen Kontexten (nicht nur im Krieg, auch oft genug an der Börse) der menschliche Akteur die Maschine nur nutzt und nicht selbst hinterfragt.

Dieses Problem ist so lange scheinbar unlösbar, wie man versucht, der KI eine explizite Kostenfunktion zu geben.

Zweites Thema: Human Compatible – wie KI gestaltet werden müsste

Stattdessen schlägt Russell vor, KI so zu gestalten, dass sie drei wesentliche Grundsätze beachtet:

1. KI soll so entwickelt werden, dass ihre einzige Zielfunktion in der bestmöglichen Realisierung menschlicher Präferenz liegt.

2. Die KI muss dabei unsicher sein, welche Handlung am besten menschliche Präferenzen realisiert – schließlich ist das niemandem, auch nicht den Menschen, in vollem Umfang klar. Die KI muss deswegen auch ihr eigenes Unwissen kennen bzw. modellieren. Und im Angesicht ihrer eigenen Unsicherheit über die menschlichen Präferenzen muss es ihr Ziel sein, diese Unsicherheit zu verringern, da sie nur so das erste Ziel erreichen kann.

3. Die maßgebliche Quelle für Informationen über menschliche Präferenzen ist das menschliche Verhalten.

Weil aber menschliche Präferenzen variabel sind (von Mensch zu Mensch, von Situation zu Situation, und über die Zeit) kann keine KI jemals vollständiges Wissen über die menschlichen Präferenzen erlangen. Stattdessen muss die KI also auch auf Dauer ihre eigenen Unsicherheiten mit modellieren – sie muss also immer mit einem statistischen oder jedenfalls unsicherheitsbehafteten Modell menschlicher Präferenz arbeiten.

Sie muss schließlich so entscheiden, dass unter dieser Unsicherheit die erwartete Übereinstimmung mit der menschlichen Präferenz am größten wird. Im Extremfall kann das auch dazu führen, dass die KI sich selbst ausschaltet, wenn das das geringste Risiko darstellt, sich gegen die Präferenzen der Menschen zu verhalten.

So erhält man, wenn die drei Prinzipien beachtet werden, endlich KI-Systeme, die auch bei herausragender „Intelligenz“ und größtmöglichem Zugriff auf Ressourcen weiter durch Menschen kontrollierbar bleiben!

Ausblick: Warum damit leider noch nicht alles klar ist

Zweifellos stellt das Konzept von Stuart Russell mindestens so viele Fragen, wie es Antworten gibt. Aus meiner Sicht ist es trotzdem das einzige Konzept, das in der Lage ist, eine Ausrichtung maschineller Intelligenz an menschlichen Zielen zu erreichen.

Wie das im Einzelnen geschieht, also, wie genau menschliche Präferenz modelliert wird, wie erreicht wird, dass das Modell auch die eigene Unsicherheit richtig einschätzt oder jedenfalls nicht unterschätzt und wie dann die vielen und oft veränderlichen und in Konkurrenz stehenden Ziele vieler Menschen zu einer bestmöglichen Handlung führen, das sind große Fragen und Aufgaben für die Entwicklung der KI in den nächsten Jahrzehnten, wenn nicht darüber hinaus.

Ich persönlich glaube aber, dass dieser Weg der einzige ist, der überhaupt zu KI-Systemen führen kann, die ihre Handlungen an humanistischen und humanen Zielen ausrichten – und damit der einzige verantwortungsvolle Weg in die Zukunft des maschinellen Lernens. Deswegen möchte ich diese Buchvorstellung mit einem Appell beenden: Lassen Sie uns Russells Strategie zu unserer eigenen machen – so wie Hilberts Programm die Mathematik auf solide Beine stellen wollte, können wir nun „Russells Strategie für menschzentrierte KI“ in den Mittelpunkt der Forschung stellen:

> *Verantwortungsvolle KI muss menschliche Ziele und Werte unterstützen.*

Es mag sein, dass das Ziel (wie schon bei Hilbert) zu hochgesteckt oder unrealistisch ist. Aber hier gibt es eine überzeugende, vielleicht die wichtigste Idee, wie eine an menschlichen Werten ausgerichtete KI vielleicht doch möglich wird – wenn wir gemeinsam und koordiniert als Forschende und als Gesellschaft in diese Richtung gehen.

Stuart Russell ist Professor für Informatik an der University of California, Berkeley. Er ist bei vielen Informatiker:innen bekannt geworden durch sein Buch *Artificial Intelligence: A Modern Approach*, das er mit Peter Norvig 1995 in der ersten Auflage veröffentlicht hat. Es wird inzwischen von weit über 1000 Universitäten weltweit verwendet und als Standardwerk der *klassischen* KI gesehen. Sein neues Buch *Human Compatible* ist im englischen Original im April 2020 und in der deutschen Übersetzung im Juni 2020 erschienen. Es setzt keine technischen Vorkenntnisse voraus und erklärt in Anhängen für Interessierte kurz die Grundlagen der modernen KI. Aus meiner Sicht möchte ich zu diesem Buch, das aktuell für 15 € als Taschenbuch erhältlich ist, meine stärkste Leseempfehlung geben: Es ist unglaublich intelligent, spannend zu lesen, hervorragend geschrieben und ganz klar und stringent argumentiert. Es zeigt gleichzeitig die Gefahren einer falsch entworfenen KI, den Stand der Diskussionen dazu, und eine für mich überzeugende, weitsichtige und fundamental richtige Strategie zur Entwicklung menschenfreundlicher Technologien.

Dorothea Kolossa

Dorothea Kolossa ist Professorin für Kognitive Signalverarbeitung an der Ruhr-Universität Bochum. In ihrer Forschung beschäftigt sie sich mit dem maschinellen Lernen aus Zeitreihen, besonders für Sprach- und Multimediadaten. In verschiedenen Projekten arbeiten sie und ihr Team an einer verantwortungsvollen und die Privatsphäre achtenden Entwicklung von Sprachtechnologien und intelligenten Systemen.