Zweitveröffentlichung

Benzmüller, Christoph

Symbolische Kontrolle für subsymbolische Künstliche Intelligenz?

Datum der Zweitveröffentlichung: 16.02.2026

Verlagsversion (Version of Record), Beitrag in Sammelwerk

Persistenter Identifikator: urn:nbn:de:bvb:473-irb-113190x

Erstveröffentlichung

Benzmüller, Christoph (2025): Symbolische Kontrolle für subsymbolische Künstliche Intelligenz?, in: Frank Schmiedchen, Alexander von Gernler, Martina Hafner, u. a. (Hrsg.), Künstliche Intelligenz und Wir : Stand, Nutzung und Herausforderungen der KI, Berlin: Springer Vieweg, S. 211–226, doi: 10.1007/978-3-662-71567-3_11

Rechtehinweis

Dieses Werk ist durch das Urheberrecht und/oder die Angabe einer Lizenz geschützt. Es steht Ihnen frei, dieses Werk auf jede Art und Weise zu nutzen, die durch die für Sie geltende Gesetzgebung zum Urheberrecht und/oder durch die Lizenz erlaubt ist. Für andere Verwendungszwecke müssen Sie die Erlaubnis der Rechteinhaberinnen und Rechteinhaber einholen.

Für dieses Dokument gilt eine Creative-Commons-Lizenz.

Die Lizenzinformationen sind online verfügbar:

https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode

Symbolische Kontrolle für subsymbolische Künstliche Intelligenz?

11

Christoph Benzmüller

Zusammenfassung

Forschungsaktivitäten in Richtung starker Künstlichen Intelligenz (KI) bzw. Artificial General Intelligence (AGI) haben aktuell stark an Fahrt aufgenommen. Ausgangspunkt sind dabei beachtliche Fortschritte vor allem im Bereich der datengetriebenen subsymbolischen KI (z. B. dem tiefen maschinellen Lernen), aber auch in der regelbasierten symbolischen KI. Mit solchen Aktivitäten im Einklang stehen sollte das Bestreben nach verantwortungsvoller KI, oder besser noch, nach sicherer KI. Selbst wenn eine starke KI (zumindest aus meiner Sicht) nicht unmittelbar absehbar ist, sollte diesem Aspekt größte Aufmerksamkeit geschenkt werden. Es ist nämlich insbesondere der unreflektierte Einsatz leistungsfähiger unvollkommener KI-Systeme in sehr kritischen Anwendungsgebieten (einschließlich des militärischen Bereichs), der zunehmend Anlass zur Sorge bereitet. In diesem Kapitel skizziere ich meine Position zu den Themen starke und sichere KI, führe dazu Äußerungen aus vorherigen eigenen Arbeiten zusammen und ergänze diese.

11.1 Einleitung

Es war die zu Beginn der 90er-Jahre vom deutschen KI-Pionier Jörg Siekmann in seinen Saarbrücker Vorlesungen in den Raum gestellte Vision einer entstehenden starken KI bzw. einer Artificial General Intelligence (AGI) und den sich daraus ergebenden Implikationen

C. Benzmüller (✉)
Otto-Friedrich-Universität Bamberg, Bamberg, Deutschland

Freie Universität Berlin, Berlin, Deutschland
E-Mail: christoph.benzmueller@uni-bamberg.de; c.benzmueller@fu-berlin.de

© Der/die Autor(en) 2026

F. Schmiedchen et al. (Hrsg.), *Künstliche Intelligenz und Wir*,
https://doi.org/10.1007/978-3-662-71567-3_11

für die Menschheit, die mich (als bis dahin weitgehend gelangweilten Informatikstudenten) provozierte und die gerade deshalb meine intrinsische Motivation für dieses Gebiet weckte (und mich deshalb auch von dem eigentlich geplanten Studienfachwechsel hin zur Sportmedizin abhielt).

Siekmann, mein späterer Doktorvater, hat uns damals auf eindrucksvolle, polarisierende Weise vermittelt, dass gerade unsere Generation Zeuge einer sehr interessanten Epoche der Geschichte werden würde, in der KI-Systeme in verschiedenen, sehr anspruchsvollen Anwendungsbereichen oder vielleicht sogar generell intelligenter als Menschen werden und dabei auch eine vom Menschen unabhängige Reproduzierbarkeit und Fortpflanzung erreichen. Er skizzierte die Entstehung einer neuen Spezies, womöglich weit intelligenter als der Mensch und auf anderen biologischen und physiologischen Grundlagen aufbauend. Diese Position attackierte mein durch eine christliche Erziehung geprägtes Weltbild, aber genau deshalb packte mich dieses Thema: Was ist der Unterschied zwischen Mensch und Maschine? Wo liegen deren individuelle Grenzen? Gibt es Kernmerkmale menschlicher Intelligenz, die eine Maschine niemals erreichen kann, und umgekehrt? Zunächst glaubte ich, Antworten auf solche Fragen im Bereich der Computermathematik, genauer des automatischen Theorembeweisens, also einem Untergebiet der regelbasierten symbolischen KI, finden zu können.

Gleichzeitig begann ich mich zunehmend für Logik und die Idee formal verifizierbarer symbolischer Modellierungen von Systemen und Theorien in Informatik, Mathematik und Philosophie zu begeistern. In meiner Diplomarbeit an der Universität des Saarlandes bei Jacques Loeckx ging es daher auch um die formale Spezifikation eines Computerprogramms aus der Medizin mit dem Ziel der formalen Verifikation von Systemeigenschaften durch interaktive oder automatische Theorembeweiser. In meiner Dissertation bei Jörg Siekmann untersuchte ich die Semantik und Automatisierbarkeit höherstufiger Logik mit besonderem Augenmerk auf die Automatisierung von Gleichheitsbeweisen. In diesen Arbeiten ging es also einerseits um Aspekte der formalen Korrektheit und Sicherheit von Computerprogrammen, die durch Theorembeweiser, also symbolische KI-Systeme, belegt werden sollten. Andererseits ging es aber auch darum, immer leistungsfähigere automatische Theorembeweiser zu entwickeln, die eines Tages sogar Mathematiker in ihren Fähigkeiten übertreffen sollten.

Mir schien, dass die beiden Aspekte – sichere (KI-)Systeme vs. leistungsfähige (KI-) Systeme – nicht wirklich im Widerspruch zueinander standen, sondern sich sogar gegenseitig beflügeln konnten. In den letzten Jahren jedoch, in denen der Begriff der KI immer mehr auf datengetriebene subsymbolische KI reduziert wurde, könnte die Kluft zwischen den beiden Aspekten kaum größer sein: Die statistischen Korrelationen subsymbolischer KI-Techniken stehen eben unglücklicherweise in starkem Widerspruch zur Idee eines beweisbar korrekten Systemverhaltens.

Ein Gegentrend hat jedoch bereits eingesetzt, da die komplementären Vor- und Nachteile beider KI-Paradigmen – symbolisch und subsymbolisch – zunehmend erkannt werden. Der nächste große KI-Hype könnte daher gerade durch eine erfolgreiche Verschmelzung dieser beiden KI-Paradigmen entstehen (bzw. beim verzögerten Druck dieses Textes

vielleicht schon entstanden sein).[1] Die Verschmelzung dieser beiden Paradigmen birgt in der Tat ein großes Potential, nicht nur in Richtung Sicherheit, sondern auch in Richtung Dateneffizienz, Nachhaltigkeit und letztlich auch in Richtung AGI.

Übrigens ist die Erkenntnis, dass zum Erreichen einer AGI gerade die Verschmelzung dieser beiden KI-Paradigmen möglicherweise notwendig ist, keineswegs neu und wird seit Jahrzehnten diskutiert. Ich erinnere mich gut an abendliche Diskussionen bei Klausuren unseres damaligen DFG-Sonderforschungsbereiches 378, *Ressourcenadaptive Kognitive Prozesse,* auf Schloss Dagstuhl, bei denen auch die sich ideal ergänzende Komplementarität dieser beiden KI-Paradigmen thematisiert wurde. Aus guten Gründen positionierten sich KI-Forschungsprojekte zu dieser Zeit jedoch typischerweise nur in einem dieser beiden Lager. Dies hatte sowohl inhaltliche als auch soziologische Gründe, denn zum einen war die technologische Entwicklung auf beiden Seiten noch nicht weit genug fortgeschritten, um sich fruchtbar einer Verschmelzung zu widmen. Zum anderen gab es eine latente soziologische Konkurrenz zwischen den beiden Lagern, wobei die symbolische Seite zunächst die Nase vorn zu haben schien. Obwohl die Idee von hybrider KI also bereits recht alt ist, gab es bisher leider mehr Konkurrenz als Kooperation zwischen den beiden Lagern.

Ich selbst publiziere seit einigen Jahren zum Thema hybride und sichere KI (s. z. B. Benzmüller & Lomfeld, 2020a, b). Versuche, bereits Ende des letzten Jahrzehnts im nationalen und Berliner Umfeld frühzeitig Forschungsprojekte und Ressourcen zu diesen Themen zu akquirieren, blieben jedoch erfolglos, was ich u. a. auch auf eine eher zögerlich wahrgenommene nationale und regionale Forschungsförderpolitik zurückführe, die weniger an wissenschaftlichem Vorsprung als an entwicklungstechnischen Aufholjagden (z. B. zur rein subsymbolischen KI) interessiert zu sein scheint, bis sich wiederum Aufholjagdsituationen zu (den dann nicht mehr so) neuen Themen ergeben.

Der folgende Text ist wie folgt gegliedert: Im Abschn. 11.2 wird der Begriff der KI kurz umrissen und eine eigene Arbeitsdefinition vorgestellt; dieser Abschnitt basiert auf Auszügen aus Benzmüller (2022), die ins Deutsche übersetzt und ergänzt wurden; s. aber auch Benzmüller und Lomfeld (2020a, b). Abschn. 11.3 diskutiert, warum eine strikte Fokussierung auf subsymbolische KI und maschinelles Lernen allein sogar als evolutionärer Rückschritt angesehen werden kann; dieser Abschnitt adaptiert Teile aus Benzmüller (2024). Abschn. 11.4 widmet sich dann der Verschmelzung von symbolischer und subsymbolischer KI, diskutiert die Vorteile einer solchen Symbiose und verweist auf laufende Arbeiten und weitere Literatur. Im Abschn. 11.5 stelle ich dann eigene Ideen und Arbeiten in Richtung sicherer KI vor; dieser Abschnitt basiert auf Auszügen aus Benzmüller (2022), die ins Deutsche übersetzt und ergänzt wurden. Ein kurzer Ausblick folgt im Abschn. 11.6.

[1] S. z. B. https://www.forbes.com/sites/danielnewman/2023/07/10/the-future-ofai-and-everything-else-is-hybrid/.

11.2 KI – Symbolisch und subsymbolisch

Die Dartmouth-Konferenz 1956 in den USA wird allgemein als die Geburtsstunde der KI angesehen, und trotz ihrer relativ jungen Geschichte hat das Gebiet bereits mehrere Winter- und Sommerperioden erlebt. Die gegenwärtige Sommerperiode scheint jedoch wesentlich intensiver und anhaltender zu sein als frühere. Tatsächlich wurden in den letzten zwei Jahrzehnten im gesamten Bereich der KI, insbesondere aber im Bereich der subsymbolischen KI (z. B. im datengetriebenen tiefen maschinellen Lernen), erhebliche Fortschritte erzielt. Dies hat in den Medien große Aufmerksamkeit erregt, und die Industrie sucht und rekrutiert händeringend KI-Experten, da die KI weithin als die Dampfmaschine des 21. Jahrhunderts angesehen wird.

Medial stark ausgeschlachtete Erfolgsgeschichten, möglicherweise auch verbunden mit wirtschaftlichen Interessen, haben dann dazu geführt, dass der Begriff KI heute in der öffentlichen Wahrnehmung weitgehend auf subsymbolische KI reduziert wurde. Gleichzeitig wird symbolische KI heute oft als *good old fashioned AI* (GOFAI) bezeichnet, was eindeutig irreführend ist, da die KI seit ihren Anfängen zwischen konnektionistischen/subsymbolischen und symbolischen Paradigmen zur Modellierung und Erklärung intelligenten Verhaltens unterscheidet und die Forschungsaktivitäten in beiden Lagern zeitlich auf diese Anfänge zurückgehen. Wie weiter unten diskutiert wird, gibt es zudem eben auch sehr gute Erfolge im Bereich der symbolischen KI, wenn auch nicht auf dem Niveau der subsymbolischen KI, die derzeit verspricht, eine sehr robuste und praktikable Wahl für viele Low-Level-Anwendungen in der Industrie zu sein.

An dieser Stelle sollen die beiden gegensätzlichen Begriffe kurz skizziert werden:

Symbolische KI Der symbolische KI-Ansatz geht davon aus, dass Intelligenz aus der Manipulation abstrakter kompositorischer und bedeutungsvoller Repräsentationen entsteht. Zu den Techniken, die in diesem Bereich verwendet werden, gehören regelbasierte Systeme und formale Logik.

Subsymbolische KI Subsymbolische KI-Ansätze realisieren intellektuelle Fähigkeiten, beispielsweise mithilfe von (tiefen) künstlichen neuronalen Netzen, d. h. Netzen von Recheneinheiten ohne semantische Bedeutung. Im Gegensatz zur Modellierung und Inferenz von Kausalitäten steht hier das Lernen statischer Zusammenhänge (z. B. zwischen Wörtern) im Vordergrund.

Beide Paradigmen haben bekannte Stärken und Schwächen, die im Folgenden diskutiert werden. Und wie bereits erwähnt, hat die Debatte, ob Intelligenz auf menschlicher Ebene durch den symbolischen oder den subsymbolischen Ansatz plausibel modelliert und erklärt werden kann, eine lange Tradition.

Die eingangs erwähnte Vision einer starken KI, d. h. einer KI, die menschliche Fähigkeiten in allen oder fast allen Bereichen übertrifft, erfordert meines Erachtens eine Verschmelzung der Techniken beider Seiten (oder eine überzeugende Erklärung, warum sich aus dem datengetriebenen subsymbolischen KI-Paradigma plötzlich und ohne weiteres Zutun symbolische KI entwickeln sollte). Für mich ist es daher gerade der Bereich der hy-

briden KI oder neurosymbolischen KI, in dem *the next big thing* zu erwarten ist. Um meinen Standpunkt besser verstehen und sehen zu können, warum ich weiterhin auf der Relevanz der symbolischen KI beharre, erscheint es mir nützlich, kurz meine persönliche Arbeitsdefinition des Begriffs KI vorzustellen:

Arbeitsdefinition KI
KI ist eine Wissenschaft von Computertechnologien, die entwickelt wurden, um intelligentes Verhalten in Maschinen zu erreichen, zu untersuchen und zu erklären. *Intelligentes Verhalten* bezieht sich dabei auf eine Reihe von Fähigkeiten, die es einer Entität ermöglichen,

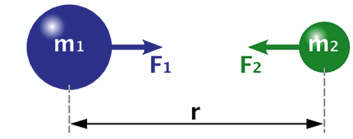
(i) bestimmte (schwierige) domänenspezifische Probleme zu lösen oder deren Lösung zu erlernen,
(ii) bekannte und unbekannte Situationen zu explorieren und zu meistern (was Wahrnehmung, Planung, Handlungsfähigkeit usw. erfordert),
(iii) rational zu denken, Widersprüche zu vermeiden und neue, abstrakte Theorien zu explorieren,
(iv) das eigene Denken und Handeln zu reflektieren und Selbstwidersprüche zu erkennen und
(v) sozial mit anderen Entitäten zu interagieren und die eigenen Ziele und Normen mit denen einer Gemeinschaft (für ein höheres Gut) in Einklang zu bringen.

Herausragende Fortschritte in der KI wurden vor allem auf Ebene (i) und in gewissem Umfang auch auf Ebene (ii) erzielt. Diese Fortschritte wurden sowohl durch subsymbolische als auch durch symbolische Techniken ermöglicht, wobei der Schwerpunkt derzeit insbesondere auf Ebene (i) eher auf den subsymbolischen Techniken zu liegen scheint. Aber auch Theorembeweiser, d. h. symbolische KI-Techniken, haben in den letzten Jahren zur Lösung schwieriger, spezifischer Probleme geführt, die der Mensch allein bisher nicht lösen konnte. Einige Beispiele aus der Mathematik finden sich in den Arbeiten von Brakensiek et al. (2022), Gonthier (2008), Gonthier et al. (2013), Hales et al. (2017), Heule (2018) und Heule und Kullmann (2017). Eigene Arbeiten adressieren die Anwendung sehr ausdrucksstarker logischer Formalismen und Theorembeweiser auch in Bereichen wie der Metaphysik und des ethisch-rechtlichen Schließens; s. Benzmüller et al. (2020) und Benzmüller und Scott (2025) sowie die weiteren Verweise darin.

Spätestens die Ebene (iii) erfordert meiner Auffassung nach die Einbeziehung symbolischer Modellierungs- und Argumentationsfähigkeiten. Insbesondere die Exploration einer neuen, abstrakten Theorie, etwa in der Mathematik, in den traditionellen Naturwissenschaften oder in der Metaphysik, setzt unweigerlich die Beherrschung einer symbolischen und tief verstandenen Repräsentationssprache voraus. Und auch das Entdecken von (z. B. versteckten, indirekten) Widersprüchen auf Ebene (iii) und (iv) kann eigentlich nur auf der Grundlage eines tiefen logisch-rationalen Verständnisses solide erfolgen.

11.3 Reine Fokussierung auf LLMs und maschinelles Lernen – ein evolutionärer Rückschritt?[2]

Insbesondere naturwissenschaftliche Erkenntnisse werden von Menschen oft in einer Mischform aus natürlicher und mathematischer Sprache dargestellt und kommuniziert, wobei auch Grafiken und Diagramme eine wichtige Rolle spielen (s. auch Abb. 11.1). Im Physikunterricht lernen Schüler zum Beispiel das Newtonsche Gravitationsgesetz kennen und führen gegebenenfalls Experimente durch, um die Plausibilität und das Erklärungspotential dieser Theorie zu hinterfragen. In der Mathematik beschäftigen sie sich zum Beispiel mit den Gesetzen der Geometrie, der Mengenlehre oder der Booleschen Algebra, sie definieren natürliche Zahlen, Quadratzahlen, Primzahlen usw. Diese symbolisch präzise definierten Begriffe werden dann als Ausgangspunkt für weitere Definitionen und Anwendungen auch über Disziplingrenzen hinweg genutzt (Primzahlen sind z. B. wichtig für Verschlüsselungsverfahren in der Kryptographie). Auf diese Weise entstehen komplexe



Kommutativgesetze	(1)	a∨b = b∨a	(1')	a∧b = b∧a
Assoziativgesetze	(2)	(a∨b)∨c = a∨(b∨c)	(2')	(a∧b)∧c = a∧(b∧c)
Absorptionsgesetze	(3)	a∧(a∨b) = a	(3')	a∨(a∧b) = a
Distributivgesetze	(4)	a∨(b∧c) = (a∨b)∧(a∨c)	(4')	a∧(b∨c) = (a∧b)∨(a∧c)
Komplementärgesetze	(5)	a∨¬a = 1	(5')	a∧¬a = 0
Neutralitätsgesetze	(6)	a∨0 = a	(6')	a∧1 = a
Extremalgesetze	(7)	a∨1 = 1	(7')	a∧0 = 0
Dualitätsgesetze	(8)	¬1 = 0	(8')	¬0 = 1
Idempotenzgesetze	(9)	a∨a = a	(9')	a∧a = a
Involutionsgesetz	(10)	¬(¬a) = a		
De Morgansche Gesetze	(11)	¬(a∨b) = ¬a∧¬b	(11')	¬(a∧b)=¬a∨¬b

Abb. 11.1 Symbolisch repräsentiertes Wissen in Physik (oben: Newtonsche Gravitationsgesetz; credit: D. Nilsson [CC BY 3.0]) und Mathematik. (Unten: Boolesche Algebra; credit: ZPG IMP [CC BY-SA 3.0 DE])

[2] Dieser Abschnitt gibt weitgehend meine Überlegungen aus Benzmüller, C. (2024): „Logikbasierte Wissensverarbeitung" wieder.

Gebäude symbolisch repräsentierten Wissens, wobei die präzise und tief verstandene Formelsprache der Mathematik oft der zentrale Informationsträger ist.

Auch in weniger mathematischen Wissenschaftszweigen gibt es wissenschaftliche Theorien, z. B. verschiedene Lerntheorien in der Didaktik und Psychologie. Hier kommt der präzisen Verwendung von natürlicher Sprache und graphischen Darstellungen eine größere Bedeutung zu. Wichtig für diesen Artikel ist die Einsicht, dass der Mensch auf der höchsten Stufe seiner erkenntnistheoretischen Schaffenskraft tief verstandene sprachliche Mittel und symbolische Darstellungen verwendet, um aus Beobachtungen oder Gedankenexperimenten abstrahiertes Wissen präzise zu beschreiben. Insbesondere die Verifikation/Falsifikation und die zielgerichtete Weiterentwicklung wissenschaftlicher Theorien werden dadurch erst ermöglicht, ebenso wie die effiziente Kommunikation von Wissen zwischen Menschen (insbesondere zur Wissensvermittlung in Schulen und Universitäten).

Der deutsche KI-Pionier Wolfgang Bibel verwendet in diesem Zusammenhang den Begriff *repräsentierende Objekte* (Bibel, 2022), also symbolische Objekte, die die Welt und ihre Eigenschaften auf einer abstrakten symbolischen Ebene repräsentieren. Besonders interessant an repräsentierenden Objekten ist, dass sie selbst zu Untersuchungsobjekten werden können; man kann sie analysieren und wiederum neues Wissen aus ihnen ableiten. Der berühmte österreichische Mathematiker Bruno Buchberger hat einen solchen Theoriefindungsprozess in Form einer Kreativitätsspirale veranschaulicht (Buchberger, 1995).

Es ist interessant, dass mit der Einführung von Computern auch Maschinen Zugang zu symbolischen, repräsentierenden Objekten erhalten haben. Dies ist der zentrale Ausgangspunkt für die symbolische KI und generell für die Informatik. Bibel fordert zu Recht, dass sich die KI viel stärker auf das Studium und die Erforschung symbolischer, repräsentierender Objekte mit Computern konzentrieren sollte. Gerade im Hinblick auf die Frage nach einer starken KI ist diese Forderung mehr als plausibel. Der aktuelle Hype um maschinelles Lernen mit neuronalen Netzen lässt diese Forderung jedoch etwas in den Hintergrund rücken. Kritisch betrachtet könnte man die aktuelle Situation daher sogar als konzeptionellen Rückschritt in Bezug auf die großen Fragen der KI betrachten, da eine starke KI ja insbesondere auch Fähigkeiten zur symbolischen Theorieexploration, z. B. in der Mathematik, besser beherrschen müsste als der Mensch.

Ein Gedankenexperiment soll diesen Einwand veranschaulichen: Nehmen wir an, wir trainieren ein neuronales Netz darauf, Primzahlen zu erkennen. Dies können wir z. B. tun, indem wir zunächst einen Datensatz mit natürlichen Zahlen, sagen wir von eins bis einer Milliarde, entsprechend annotieren (Primzahlen werden z. B. grün markiert, alle anderen Zahlen rot). Dann füttern wir ein Lernverfahren mit dieser annotierten Information und erhalten einen leistungsfähigen Primzahl-Klassifikator, den wir im Folgenden PP (für Primzahl-Papagei) nennen.

Es ist zu erwarten, dass PP sehr leistungsfähig und sehr schnell sein wird. PP wäre mir sicherlich überlegen, insbesondere bis zur Zahlengrenze von einer Milliarde, da er diese Primzahlen ja gewissermaßen *auswendig gelernt* hat (in einem ressourcenaufwendigen

Trainingsprozess und *ohne dabei zu verstehen*). Auch jenseits dieser Zahlengrenze wäre PP im *Erraten* von Primzahlen unter Zeitdruck und ohne Hilfsmittel normalen Menschen möglicherweise mindestens ebenbürtig.

Der Leser erkennt vermutlich schon, worauf dieses Gedankenexperiment hinauswill: Trotz der anzunehmenden Stärke von PP (gegenüber den meisten Menschen) bei der Erkennung von Primzahlen gibt es doch einen fundamentalen Unterschied zur menschlichen Erkenntnis: Der Primzahl-Papagei PP hätte nicht die geringste Ahnung davon, was eine Primzahl eigentlich ist, d. h. wie sie sich in ihren zahlentheoretischen Eigenschaften von anderen Zahlen unterscheidet und zu ihnen in Beziehung steht. Das ist aber auch nicht verwunderlich, da wir PP in unserem Gedankenexperiment ja auch keine entsprechenden Hintergrundinformationen mitgegeben haben. Die Frage an PP, warum er z. B. die Zahl 999.999.937 korrekt als Primzahl klassifiziert, wäre also sinnlos. PP könnte bestenfalls antworten: *„Weil mir dies in meiner Trainingsphase so mitgeteilt wurde“*. Und wenn PP z. B. bei der Klassifikation von 1.999.999.927 (eine Primzahl größer als eine Milliarde) scheitert, so könnte er uns dieses Scheitern ebenso wenig erklären. Ein symbolisch beschriebener Primzahlalgorithmus hingegen könnte beim Auftreten solcher Fehler in der Anwendung genau analysiert und der Fehler in der symbolischen Beschreibung lokalisiert und entsprechend korrigiert werden. Bei einem solchen Vorgehen betrachten wir also die symbolische Repräsentation eines Primzahltests selbst als Untersuchungsgegenstand, den wir reflektieren, analysieren, testen und ggf. zielgerichtet korrigieren wollen. Das antrainierte, undurchsichtige Modell von PP bietet dafür hingegen keine sinnvolle Grundlage, es ist einer solchen Analyse nicht zugänglich.

Aus erkenntnistheoretischer Sicht und aus Sicht der starken KI muss der Primzahl-Papagei PP aus unserem Gedankenexperiment daher – trotz seiner praktischen Stärke – als eher uninteressant eingestuft werden. Und es gibt weitere, bereits angedeutete Nachteile der subsymbolischen KI, die oft nicht genügend beachtet werden: Während erlernte Modelle bei notwendigen Anpassungen in der Regel immer wieder neu- oder nachtrainiert werden müssen, genügen zur Korrektur bzw. Anpassung explizit repräsentierten Wissens in der symbolischen KI einzelne Eingriffe/Modifikationen an der richtigen Stelle; diese korrigierten Repräsentationen können dann elegant und effizient an andere Menschen kommuniziert werden.

Auch sollten wir uns bewusst machen, dass wissenschaftliche Theorien nicht selten auf der Basis reiner Gedankenexperimente entwickelt werden. Es gibt dann keine verfügbaren Daten, die als Ausgangspunkt für das Training eines Modells herangezogen könnten.

Ein prominentes, polarisierendes Beispiel aus dem Bereich der Metaphysik ist Anselm von Canterburys Gedankenexperiment zur Existenz Gottes, auch bekannt als ontologischer Gottesbeweis oder ontologisches Argument, das seit einem Jahrtausend Gegenstand philosophischer und theologischer Untersuchungen ist; siehe z. B. die Diskussion in Benzmüller (2022). Wie könnte eine rein datenbasierte, subsymbolische KI in einem solchen Beispielkontext sinnvoll und gewinnbringend eingesetzt werden? Wie könnte sie ein solches Argument auf der Grundlage eines Gedankenexperiments selbst entwickeln?

Zu den verschiedenen Varianten des ontologischen Arguments von Anselm, die seither entwickelt wurden, gehört auch das von Kurt Gödel in der modernen Modallogik

formulierte und später von Dana Scott und anderen modifizierte *modale ontologische Argument*, mit dem ich mich im letzten Jahrzehnt zusammen mit Kollegen intensiv beschäftigt habe; s. Benzmüller und Scott (2025) und die weiteren Verweise darin. Diese Forschung, die in fruchtbarer Mensch-Maschine-Interaktion unter Verwendung symbolischer Repräsentations- und Kommunikationstechniken durchgeführt wurde, hat unter anderem zur computergestützten Exploration alternativer und zum Teil stark vereinfachter Varianten geführt (s. z. B. Benzmüller, 2022), die wiederum das menschliche Verständnis des ontologischen Arguments fördern können. Während bisher nur Teile des Explorationsprozesses durch symbolische KI-Systeme, konkret Theorembeweiser und Modellgenerierer, automatisch unterstützt werden, sehe ich zunehmend Anzeichen dafür, dass der Automatisierungsgrad solcher Aktivitäten in Zukunft (besonders wenn hybride KI-Systeme eingesetzt werden) deutlich erhöht werden kann. In rein subsymbolischen KI-Systemen scheinen jedoch weder die hier erreichte abstrakte symbolische Theorieexploration noch die anschließende Theorieevaluation oder eine sinnvolle symbolische Kommunikation mit dem Menschen hinreichend gut unterstützt zu sein, um eine solche fruchtbare Mensch-Maschine-Interaktion im Bereich der Metaphysik abzubilden.

11.4 Die Zusammenführung der beiden konkurrierenden KI-Paradigmen

Während die Vorteile und die Notwendigkeit der Integration von symbolischer und subsymbolischer KI von vielen Wissenschaftlern schon seit geraumer Zeit betont werden (für aktuelle Texte s. z. B. Lenat & Marcus, 2023; Marcus, 2023b; Marra et al., 2024; Pantsar, 2024; Platzer, 2024; Tao, 2024), gibt es in jüngster Zeit aber auch erste beeindruckende praktische Erfolge. So hat das System *AlphaGeometry* von DeepMind (Trinh et al., 2024), das maschinelles Lernen und automatisches Theorembeweisen auf innovative Weise kombiniert, um unter anderem zahlreiche mathematisch abgesicherte Trainingsdaten für Geometriebeweise zu synthetisieren, erstaunliche Erfolge bei der Lösung von Geometrieaufgaben der Mathematik-Olympiade erzielt. Aber auch bei der Integration von Techniken des maschinellen Lernens in beispielsweise interaktive oder automatische Theorembeweiser gab es im letzten Jahrzehnt signifikante Fortschritte. Beispiele werden diskutiert von Fulton und Platzer (2018), Schulz und Möhrmann (2016) sowie Olsák et al. (2020); weitere Verweise finden sich in den Übersichten von Blaauwbroek et al. (2024) sowie auch Platzer (2024).

Auch die Idee der Autoformalisierung mathematischer Beweise (Szegedy, 2020; Wu et al., 2022), also die Idee, robuste Transformationen natürlichsprachlicher mathematischer Texte und Beweise in formalen Repräsentationen zu erlernen, sollte in diesem Zusammenhang erwähnt werden, da gerade dieser Ansatz im Erfolgsfall ein großes Potential für sehr innovative Verschmelzungen mit Rückkopplungen zwischen beiden KI-Paradigmen liefern kann.

Die Komplementarität der Vorteile von symbolischer und subsymbolischer KI ist in Abb. 11.2 kurz skizziert. Als Nachteile beider Paradigmen wurden u. a. folgende Aspekte identifiziert:

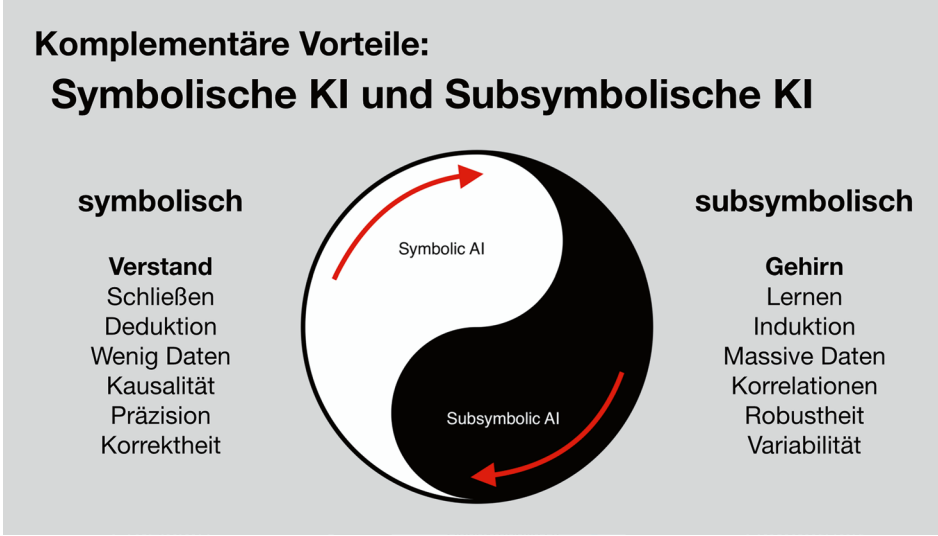


Abb. 11.2 Komplementäre Vorteile der beiden KI-Paradigmen

Subsymbolische KI

Halluzinationen KI-Sprachmodelle haben keinen direkten Zugang zu Weltwissen, und die Antworten und Inhalte, die diese generieren, basieren auf der statistischen Nähe von z. B. Wörtern in Trainingsdaten. Insbesondere bei neuartigen Anfragen kann es deshalb leicht zu Halluzinationen kommen, d. h. frei erfundenen Antworten, welche auf typischen Wortreihenfolgen basieren.[3]

Intransparenz Bei KI-Modellen ist es in der Regel nicht oder kaum möglich, kausale Entscheidungs- und Berechnungsgrundlagen zu hinterfragen, weil die Berechnungen dieser Systeme primär auf Korrelationen und statistischer Nähe beruhen und zudem sehr komplexe Strukturen aufweisen (zur Erinnerung: der Primzahl-Papagei PP aus dem vorherigen Kapitel würde vielleicht richtig erraten, dass 999.999.937 eine Primzahl ist, aber er könnte uns nicht sagen, warum das so ist). Da also wie bei einem Orakel, bzw. einer Black-Box, Informationen ohne belastbare Begründungen zurückgegeben werden, entstehen ethische und rechtliche Probleme, die auch durch die aktuellen Techniken im Bereich der erklärbaren KI nur teilweise adressiert werden können.

Datenhunger Für das Training von KI-Modellen werden in der Regel sehr große Datenmengen und Rechenressourcen benötigt, so dass subsymbolische KI-Technologien selbst zu einem großen Nachhaltigkeitsproblem werden. Auch die mangelnde Qualität großer Datenmengen, z. B. im Internet (die zunehmend durch neue KI-synthetisierte Daten angereichert und ggf. *verschmutzt* werden), stellt ein Problem dar, da mangelnde Datenqualität zu mangelhaften KI-Systemen führt. Deshalb erscheint es in vielen Anwendungsbereichen geradezu absurd, von Menschen entwickelte symbolische Theorien

[3] S. dazu auch den Blogbeitrag „How come GPT can seem so brilliant one minute and so breathtakingly dumb the next?“ von Gary Marcus unter http://bit.ly/3wL4Ir4.

und Algorithmen (wie z. B. einen Primzahltestalgorithmus in der Mathematik) in große Datenmengen (zurück) zu übersetzen, um dann aus diesen Daten wiederum ein intransparentes KI-Modell mit enormem Rechenaufwand zu trainieren. Ich persönlich sehe dies als einen evolutionären Rückschritt an, der gleichzeitig ethische Fragen aufwirft.

Symbolische KI

Fragilität Regel-, logik- und wissensbasierte Systeme in der symbolischen KI sind von Natur aus interpretierbar, transparent und nachvollziehbar, aber leider auch recht fragil. Damit ist die relative Instabilität solcher Systeme gegenüber Änderungen und Anpassungen in der Anwendungsdomäne gemeint. Notwendige Anpassungen können oft nur von Experten vorgenommen werden.

Domänenspezifität Symbolische KI-Systeme werden typischerweise für klar definierte und relativ enge Anwendungsdomänen entwickelt. Das Anwendungsspektrum solcher Systeme ist daher nur schwer erweiterbar bzw. verallgemeinerbar. Besonders in sehr komplexen und dynamischen Anwendungsbereichen sind ihre Einsatzfähigkeit und Relevanz daher oft sehr eingeschränkt.

Schlechte Skalierbarkeit Die Entwicklung symbolischer KI-Systeme und -Algorithmen erfordert oft einen hohen manuellen Aufwand durch Domänenexperten. In komplexen und dynamischen Anwendungsdomänen ist dieser manuelle Aufwand für die Erstellung symbolischer Modellierungen jedoch oft zu hoch oder nahezu unmöglich. Hinzu kommen Performanzprobleme symbolischer KI-Algorithmen, z. B. bei der Suche in großen Datenmengen oder Lösungsräumen.

Forderungen nach einer Zusammenführung bzw. Integration von Techniken der symbolischen und subsymbolischen KI sind aufgrund der skizzierten Komplementarität der Vor- und Nachteile naheliegend und vielversprechend. Hinsichtlich des methodischen Vorgehens bei einer solchen Zusammenführung unterscheide ich persönlich wie folgt zwischen den Begriffen *hybride KI* und *neurosymbolische KI*:

Hybride KI Hybride KI bezeichnet die Zusammenführung und Verknüpfung von Techniken und Methoden aus den Bereichen der subsymbolischen und der symbolischen KI mit dem Ziel, durch die Komplementarität der Schwächen und Stärken beider Ansätze insgesamt bessere und sicherere KI-Systeme zu schaffen. Eine Verschmelzung beider Ansätze ist dabei nicht zwingend.

Neurosymbolische KI Unter neurosymbolischer KI verstehe ich eine hybride KI, bei der symbolische Fähigkeiten z. B. direkt auf der Ebene eines neuronalen Netzes realisiert sind. Hier steht also die konzeptuelle Verschmelzung beider Ansätze in einem einzigen homogenen, kohärenten Gesamtmodell/-system stärker im Vordergrund.

11.5 Symbolische Schutzhüllen für subsymbolische KI-Systeme?

In verschiedenen Vorträgen und Texten habe ich mich frühzeitig für die Entwicklung symbolischer Kontrollmechanismen für (subsymbolische) KI-Systeme ausgesprochen, um insbesondere den in Abschn. 11.1 beschriebenen Spagat zwischen Verifizierbarkeit und

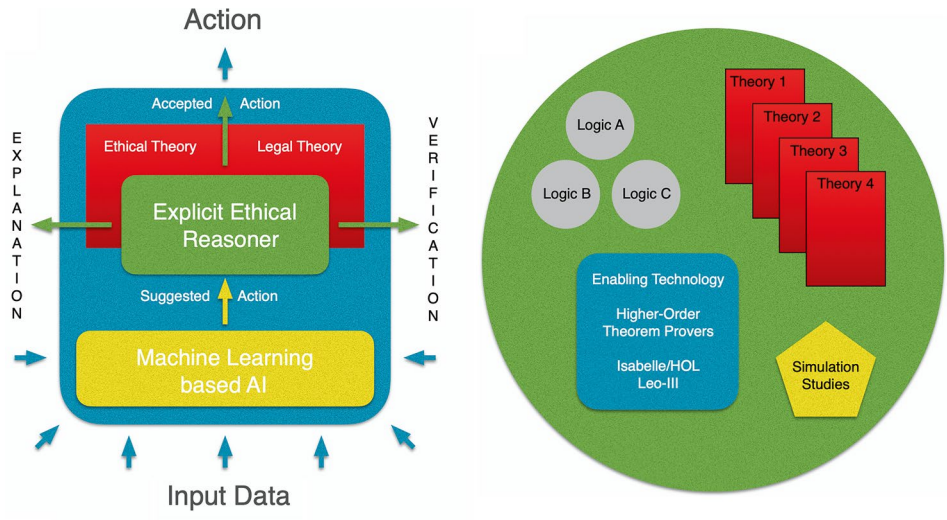


Abb. 11.3 Vorschlag einer explizit-ethischen Kontrolle für (subsymbolische) KI-Systeme (links); die erforderliche explizite Modellierung verschiedener ethisch-rechtlicher Theorien setzt jedoch die Exploration geeigneter Logikformalismen voraus (rechts); s. dazu auch Benzmüller et al. (2020) und die weiteren Verweise darin

Leistungsfähigkeit moderner subsymbolischer KI-Systeme zu adressieren (Benzmüller & Lomfeld, 2020a, b; Benzmüller et al., 2020). Diese Arbeiten sind motiviert durch meine Überzeugung, dass gerade die hybride KI gute Chancen bietet, verifizierbar sichere KI-Systeme zu entwickeln.

Betrachten wir dazu den in Abb. 11.3 links dargestellten Vorschlag eines explizit ethischen KI-Agenten. Wie in einem vorherigen Artikel (Benzmüller et al., 2020) ausführlicher diskutiert, wird in der dargestellten Architektur ein (z. B. subsymbolisches) KI-System (in Gelb, z. B. ein intelligentes autonomes System zur Entscheidungsfindung in einer bestimmten Domäne) zu einem explizit ethischen Agenten angereichert, indem dem gelben System eine symbolische Kontroll- und Reflexionskomponente hinzugefügt wird, welche die vom diesem System vorgeschlagenen Handlungsoptionen einer nachgeschalteten ethisch-rechtlichen Bewertung unterzieht. Diese nachgelagerte Bewertung soll eine zusätzliche, explizite Ebene der ethisch-rechtlichen Kontrolle über den unberechenbaren und schlecht verifizierbaren KI-Agenten in Gelb realisieren.

Die Details der Architektur des gelben KI-Systems sind hier nicht von Bedeutung. Beispielsweise ist es unerheblich, ob seine Berechnungen auf subsymbolischen oder symbolischen Techniken oder Kombinationen davon beruhen. Wichtig ist jedoch, dass die vom gelben KI-Agenten vorgeschlagenen Aktionen – zumindest die als besonders kritisch eingestuften – nicht unmittelbar ausgeführt werden. Stattdessen werden sie vor einer möglichen Ausführung zusätzlich auf ihre Vereinbarkeit mit einer explizit modellierten ethisch-rechtlichen Theorie geprüft. Sicherheitsrelevant ist also weniger die korrekte Funktionsweise des eingekapselten gelben KI-Agenten, sondern die (idealerweise verifizierbare)

korrekte Funktionsweise der symbolischen Schutzhülle. Insbesondere in hochkritischen Kontexten kommt es eben viel weniger darauf an, wie ein KI-System seine präferierten Handlungsoptionen bestimmt, sondern vor allem darauf, ob diese Handlungsoptionen die zusätzliche Bewertung durch eine übergeordnete explizit-ethische Schutzhülle bestehen, bevor sie ausgeführt werden dürfen. Es ist zu beachten, dass es in einigen kritischen Anwendungskontexten von besonderem Interesse ist, die Offenlegung der genauen Funktionsweise des internen gelben Systems gerade zu vermeiden, um Risiken (z. B. gezielte Manipulation) zu minimieren.

Das zusätzliche Yin-Yang-Symbol in der Abb. 11.2 links weist zudem darauf hin, dass eine solche Architektur auch ein großes Potential für die Entwicklung von Techniken zur *Anpassung* bzw. *Aushandlung* von bottom-up erlerntem normativem Verhalten gegenüber top-down postulierten normativen Vorgaben bietet. Generell ergeben sich hier auch interessante Fragestellungen in Richtung AGI, und die Parallelen zu den von Kahneman (2011) in seinem Bestseller *„Thinking – Fast and Slow"* skizzierten Systemebenen 1 und 2 sind recht offensichtlich.

Damit die symbolische Schutzhülle mit ethisch-rechtlichen Domänentheorien adäquat bestückt werden kann, sind typischerweise sehr ausdrucksstarke Logiken und Logikkombinationen als Repräsentationsformalismen erforderlich; s. auch Abb. 11.3 (rechts). Die Entwicklung solcher Logikformalismen (z. B. für normatives Schließen) ist jedoch derzeit selbst noch Gegenstand wissenschaftlicher Forschung, die aber ihrerseits wieder durch symbolische KI-Systeme unterstützt werden kann.

Insbesondere die adäquate und korrekte Modellierung normativer Konzepte, wie Erlaubnis und Verbot, ist in der Praxis oft komplexer und schwieriger, als man auf den ersten Blick vermuten würde. Dies zeigt sich unter anderem in den berüchtigten Paradoxien des normativen Schließens, und adäquate Lösungen für diese Herausforderung erfordern oft die Verwendung anspruchsvoller deontischer Logiken (Gabbay et al., 2013).

In unserem logikpluralistischen Wissensrepräsentationsansatz LogiKEy (Benzmüller et al., 2020) modellieren und implementieren wir daher ethisch-rechtliche Theorien, normatives Schließen und deontische Logiken letztendlich in einer ausdrucksstarken höherstufigen Metalogik, die als einziger Fixpunkt in dem ansonsten logikpluralistischen Rahmen fungiert und für die in den letzten Jahren immer leistungsfähigere Theorembeweiser entwickelt wurden, die nun als universelle Logikmaschinen im LogiKEy-Kontext eingesetzt werden können.

11.6 Diskussion und Ausblick

Explizite, symbolische Repräsentationen von ethisch-rechtlichen Vorgaben sind für die Realisierung vertrauenswürdiger und sicherer KI-Systeme besonders relevant, weil sie normatives (und anderes) Wissen nicht nur transparent und erklärbar, sondern auch effizient und robust zwischen Mensch und Maschine kommunizierbar machen.

Zusammen mit Lomfeld habe ich deshalb für hybride Lösungen plädiert, *„die intelligente Systeme dazu befähigen, „echte Gründe" für ihre Handlungen und Entscheidungen*

zu generieren und zu kommunizieren (Benzmüller & Lomfeld, 2020a). *Während „klassische" Ansätze interpretierbarer KI nach transparenten Erklärungen für (subsymbolische) Maschinenprozesse suchen, möchten wir „Vertrauenswürdigkeit durch rationale Kommunikation" erreichen, d. h. Maschinen sollen in realer sozialer Interaktion Gründe für ihr Handeln austauschen. Dieser interdisziplinäre Forschungsvorschlag führt innovative Ansätze aus symbolischer KI, maschinellem Lernen, Mensch-Maschinen-Interaktion, Recht und Philosophie zusammen"* (Benzmüller & Lomfeld, 2020b).

Explizit verfügbare symbolische Rechtfertigungen können dann zusätzlich als Ansatzpunkte für normative, symbolische Kontroll- und Steuerungskomponenten dienen, wie sie z. B. im vorangegangenen Abschnitt skizziert wurden. Die Entwicklung solcher sicheren KI-Systeme erfordert aber vor allem ausreichend Zeit und Ressourcen, und beides bereitzustellen halte ich im gegenwärtigen globalen Klima für eine große (forschungs-)politische Herausforderung.

Ich habe Forderungen nach einer sinnvollen Regulierung, wie sie beispielsweise von Marcus (2023a) diskutiert werden, stets unterstützt und mich dafür ausgesprochen, in besonders kritischen Anwendungsbereichen (insbesondere im Hinblick auf eine übereilte und unreflektierte Entwicklung und Nutzung von KI im militärischen Kontext) nur mit größter Vorsicht vorzugehen. Während der europäische *AI Act* (European Commission, 2021; High-Level Expert Group on AI, 2020) mit seinem Stufenmodell einen guten Einstieg in eine sinnvolle Regulierung darstellt, kommt es nun auf eine adäquate Umsetzung, internationale Ausdehnung und konsequente Fortführung dieser eingeschlagenen Richtung an. Eine besondere Herausforderung stellt dabei der Umgang mit kritischen Anwendungsbereichen von KI-Systemen dar, weil diese zum Teil sehr unterschiedliche Ansätze und Lösungen erfordern werden. Gerade in solchen Domänen halte ich symbolische Schutzhüllen für subsymbolische KI-Systeme für einen sehr vielversprechenden Ansatz in Richtung sicherer KI-Systeme.

Literatur

Benzmüller, C. (2022). Symbolic AI and Gödel's ontological argument. *Zygon(r), 57*(4), 953–962. https://doi.org/10.1111/zygo.12830

Benzmüller, C. (2024). Logikbasierte Wissensverarbeitung. In U. Furbach et al. (Hrsg.), *Künstliche Intelligenz für Lehrkräfte – Eine fachliche Einführung mit didaktischen Hinweisen* (S. 139–162). Springer Vieweg. https://doi.org/10.1007/978-3658-44248-4_11

Benzmüller, C., & Lomfeld, B. (2020a). Reasonable Machines: A Research Manifesto. In U. Schmid et al. (Hrsg.), *KI 2020: Advances in Artificial Intelligence – 43rd German Conference on Artificial Intelligence, Proceedings. Bamberg, Deutschland. 21.–25. September 2020* (Lecture Notes in Artificial Intelligence, Bd. 12352, S. 251–258. ISBN: 978-3-030-30178-1). Springer. https://doi.org/10.1007/978-3-03058285-2_20

Benzmüller, C., & Lomfeld, B. (2020b). Träumen vernünftige Maschinen von Gründen? Eine reale Utopie. In S. Ammon et al. (Hrsg.), *Verantwortung KI – Künstliche Intelligenz und gesellschaftliche Folgen (Berlin-Brandenburgische Akademie der Wissenschaften)*. https://www.bbaw.de/fi-

lesbbaw/user_upload/publikationen/BBAW_Verantwortung-KI-3-2020_PDF-A-1b.pdf. Zugegriffen am 18.07.2025.

Benzmüller, C., & Scott, D. S. (2025). Notes on Gödel's and Scott's variants of the ontological argument. *Monatshefte für Mathematik*. In Print. https://doi.org/10.1007/s00605-025-02078-x

Benzmüller, C., et al. (2020). Designing normative theories for ethical and legal reasoning: LogiKEy framework, methodology, and tool support. *Artificial Intelligence, 287*, 103348. https://doi.org/10.1016/j.artint.2020.103348. ISSN: 0004-3702.

Bibel, W. (2022). Komputer kreiert Wissenschaft. *Informatik Spektrum, 45*(6), 356–365. https://doi.org/10.1007/S00287-022-01456-1

Blaauwbroek, L., et al. (2024). Learning guided automated reasoning: A brief survey. In V. Capretta et al. (Hrsg.), *Logics and type systems in theory and practice – Essays dedicated to herman geuvers on the occasion of his 60th birthday* (Lecture notes in computer science, Bd. 14560, S. 54–83). Springer. https://doi.org/10.1007/978-3-031-61716-4_4

Brakensiek, J., et al. (2022). The resolution of Keller's conjecture. *Journal of Automated Reasoning, 66*(3), 277–300. https://doi.org/10.1007/s10817-022-09623-5

Buchberger, B. (1995). https://www.researchgate.net/figure/Mathematical-Creativity-Spiral-Buchberger-1995_fig4_221562312. Zugegriffen am 18.07.2025.

European Commission. (2021). *The AI Act, COM(2021)206 final*. https://artificialintelligenceact.eu/the-act/. Zugegriffen am 18.07.2025.

Fulton, N., & Platzer, A. (2018). Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In S. McIlraith, & K. Q Weinberger (Hrsg.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18). New Orleans, Louisiana, USA, 02.–07. Februar 2018* (S. 6485–6492). AAAI Press. https://doi.org/10.1609/AAAI.V32I1.12107

Gabbay, D., et al. (2013). *Handbook of deontic logic and normative systems* (Bd. 1). College Publications.

Gonthier, G. (2008). Formal proof – The four-color theorem. In *Notices of the American Mathematical Society* (Bd. 55.11, S. 1382–1393) http://www.ams.org/notices/200811/tx081101382p.pdf

Gonthier, G., et al. (2013). A machine-checked proof of the odd order theorem. In S. Blazy et al. (Hrsg.), *Interactive theorem proving* (Lecture notes in computer science, Bd. 7998, S. 163–179). Springer. https://doi.org/10.1007/978-3-642-396342_14

Hales, T., et al. (2017). A formal proof of the Kepler conjecture. *Forum of Mathematics, Pi, 5*, e2. https://doi.org/10.1017/fmp.2017.1

Heule, M. J. H. (2018). Schur Number Five. In *AAAI* (S. 6598–6606). AAAI Press. https://dl.acm.org/doi/pdf/10.5555/3504035.3504843

Heule, M. J. H., & Kullmann, O. (2017). The science of brute force. *Communications of the ACM, 60*(8), 70–79. https://doi.org/10.1145/3107239. ISSN: 0001-0782.

High-Level Expert Group on AI. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment*. https://digitalstrategy.ec.europa.eu/en/library/assessment-list-trustworthyartificial-intelligence-altai-self-assessment. Zugegriffen am 18.07.2025.

Kahneman, D. (2011). *Thinking, fast and slow*. Allen Lane.

Lenat, D., & Marcus, G. (2023). Getting from generative AI to trustworthy AI: What LLMs might learn from Cyc. *arXiv*, 2308.04445. https://doi.org/10.48550/arXiv.2308.04445

Marcus, G. (2023a). Controlling AI. *Communications of the ACM, 66*(10), 6–7. https://doi.org/10.1145/3613250

Marcus, G. (2023b). Hoping for the best as AI evolves. *Communications of the ACM, 66*(4), 6–7. https://doi.org/10.1145/3583078

Marra, G., et al. (2024). From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial intelligence, 328*, 104062. https://doi.org/10.1016/J.ARTINT.2023

Olsák, M., et al. (2020). Property Invariant Embedding for Automated Reasoning. In G. De Giacomo, et al. (Hrsg.), *ECAI 2020 – 24th European Conference on Artificial Intelligence, Santiago de Compostela, Spanien, 29. August – 08. September 2020. Incl. 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)* (Frontiers in artificial intelligence and applications, Bd. 325, S. 1395–1402). IOS Press. https://doi.org/10.3233/FAIA200244.

Pantsar, M. (2024). Theorem proving in artificial neural networks: New frontiers in mathematical AI. *European Journal of Philosophy of Science, 14*, 4. https://doi.org/10.1007/s13194-02400569-6

Platzer, A. (2024). Intersymbolic AI: Interlinking symbolic AI and subsymbolic AI. *CoRR abs/2406.11563, arXiv*, 2406.11563. https://doi.org/10.48550/ARXIV.2406.11563

Schulz, S., & Möhrmann, M. (2016). Performance of clause selection heuristics for saturation-based theorem proving. In N. Olivetti & A. Tiwari (Hrsg.), *Automated reasoning. 8th international joint conference, IJCAR 2016, Coimbra, Portugal, 27. Juni – 02. Juli 2016, Proceedings* (Lecture notes in computer science, Bd. 9706, S. 330–345). Springer. https://doi.org/10.1007/978-3-319-40229-1_23

Szegedy, C. (2020). A Promising Path Towards Autoformalization and General Artificial Intelligence. In C. Benzmüller & B. R. Miller (Hrsg.), *Intelligent computer mathematics 13th international conference, CICM 2020, Proceedings. Bertinoro, Italien, 26.–31. Juli 2020* (Lecture notes in computer science, Bd. 12236, S. 3–20). Springer. https://doi.org/10.1007/978-3-030-53518-6_1

Tao, T. (2024). *Machine assisted proof*. https://terrytao.wordpress.com/wp-content/uploads/2024/03/machine-assisted-proof-notices.pdf. Zugegriffen am 18.07.2025.

Trinh, T. H., et al. (2024). Solving olympiad geometry without human demonstrations. *Nature, 625*(7995), 476–482. https://doi.org/10.1038/S41586-023-06747-5

Wu, Y., et al. (2022). Autoformalization with large language models. In S. Koyejo et al. (Hrsg.), *Advances in neural information processing systems 35: Annual conference on neural information processing systems 2022, NeurIPS 2022, New Orleans, LA, USA, 28. November – 09. Dezember 2022*. https://proceedings.neurips.cc/paper_files/paper/2022/file/d0c6bc641a56bebee9d985b937307367-PaperConference.pdf. Zugegriffen am 18.07.2025.

Open Access Dieses Kapitel wird unter der Creative Commons Namensnennung - Nicht kommerziell - Keine Bearbeitung 4.0 International Lizenz (http://creativecommons.org/licenses/by-nc-nd/4.0/deed.de) veröffentlicht, welche die nicht-kommerzielle Nutzung, Vervielfältigung, Verbreitung und Wiedergabe in jeglichem Medium und Format erlaubt, sofern Sie den/die ursprünglichen Autor(en) und die Quelle ordnungsgemäß nennen, einen Link zur Creative Commons Lizenz beifügen und angeben, ob Änderungen vorgenommen wurden. Die Lizenz gibt Ihnen nicht das Recht, bearbeitete oder sonst wie umgestaltete Fassungen dieses Werkes zu verbreiten oder öffentlich wiederzugeben.

Die in diesem Kapitel enthaltenen Bilder und sonstiges Drittmaterial unterliegen ebenfalls der genannten Creative Commons Lizenz, sofern sich aus der Abbildungslegende nichts anderes ergibt. Sofern das betreffende Material nicht unter der genannten Creative Commons Lizenz steht und die betreffende Handlung nicht nach gesetzlichen Vorschriften erlaubt ist, ist auch für die oben aufgeführten nicht-kommerziellen Weiterverwendungen des Materials die Einwilligung des jeweiligen Rechteinhabers einzuholen.