MARSILIUS KOLLEG

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

MENSCH UND AUTOMAT – DIE ROLLE VON ZUFALL UND DETERMINISMUS

Fellowbericht

Ullrich Köthe

DOI: 10.11588/fmk.2022.2.92716

MARSILIUS-KOLLEG
2021 / 2022

ULLRICH KÖTHE

MENSCH UND AUTOMAT

Die Rolle von Zufall und Determinismus

Mit der stürmischen Entwicklung der Automaten durch die aktuellen Fortschritte in der künstlichen Intelligenz ist die Frage nach dem Verhältnis zwischen Menschen und Automaten wichtiger denn je. Das Marsilius-Projekt unseres Teams Rebecca Müller, Andreas Voß und Ullrich Köthe hat sich dieser Frage von drei verschiedenen Seiten genähert: der Kunstgeschichte, der Psychologie und dem maschinellen Lernen. Auf diese Weise konnten wir vergleichen, wie Menschen der Vergangenheit sich wirklichen oder mythischen Automaten gegenüber verhalten haben, wie dieses Verhältnis aus psychologischer Sicht heute gestaltet werden sollte und welche Möglichkeiten das maschinelle Lernen für diese Gestaltung bietet bzw. welche Beschränkungen es (noch) gibt.

BEDEUTUNG DES ZUFALLS

In meinem eigenen Projektteil habe ich mich besonders auf die Probleme der Unsicherheit und des Zufalls fokussiert. Das Verblüffende am Phänomen „Zufall" sind seine komplementären Effekte: einerseits erschweren nicht-deterministische Vorgänge in der Natur (zufällige Messfehler usw.) das genaue Verständnis der Realität und ihre Vorhersage. Andererseits sind Zufallsprozesse grundlegend für die Ausprägung von Vielfalt und Variabilität in der Natur, der Gesellschaft und auch der künstlichen Intelligenz. Für das maschinelle Lernen hat diese Komplementarität eine paradoxe Konsequenz, die ich im Rahmen dieses Projektes überzeugend darlegen konnte: Um Probleme und Schwierigkeiten zu überwinden, denen Menschen und Automaten durch die Unsicherheit und Zufälligkeit der Welt unterworfen sind, hilft es, in die Analyse- und Entscheidungsverfahren *zusätzliche* Zufallskomponenten – in Form randomisierter Algorithmen – einzubauen. In gewissem Sinne kann man hier tatsächlich den Teufel mit dem Beelzebub austreiben!

Ein wichtiges Ergebnis meines Projekts war es, die Quellen der Unsicherheit unserer Beschreibungen der Welt genauer zu klassifizieren. Erstens kann die Welt inhärente Beschränkungen der Vorhersagbarkeit haben – etwa bei Phänomenen der Quantenphysik, deren beobachtbare Effekte nur probabilistisch vorhergesagt werden können. Zweitens sind unsere Beobachtungen über die Welt stets unvollkommen und unter anderem durch das Rauschen der Messgeräte verfälscht. Drittens schließlich sind unsere Beschreibungen der Welt (unsere „mentalen Modelle") unsicher, weil kein Modell die wahre Komplexität der Welt vollständig wiedergeben kann.

Besonders interessant ist, dass die Unsicherheit auf allen drei Ebenen nicht nur durch Zufallsprozesse, sondern auch durch hochkomplexe deterministische Zusammenhänge verursacht werden kann. Bei der inhärenten Unsicherheit der Welt wird dies z. B. durch „deterministisches Chaos" verursacht, das beispielsweise eine Wettervorhersage über einen längeren Zeitraum als zwei Wochen selbst dann ausschlösse, wenn die zugrundeliegenden Informationen und Modelle perfekt wären. Auch innerhalb der Beobachtungen tritt oft ein deterministischer Informationsverlust auf, wie bei der perspektivischen Projektion der dreidimensionalen Welt auf zweidimensionale Detektoren (Retina oder Kamera). Auf der Ebene der Modelle entstehen z. B. deterministische Fehler, wenn sich das Verhalten der Welt gegenüber dem Zeitpunkt der Modelldefinition geändert hat, wie es beispielsweise bei Mutationen des SARS-CoV-2-Virus passiert, die dessen Gefährlichkeit und Übertragbarkeit verändern und damit bisherige Modelle invalidieren.

Daraus folgt, dass eine (die?) grundlegende Fähigkeit von Menschen und Automaten im Umgang mit der Welt darin besteht, Entscheidungen über zielführende Aktionen treffen zu können, *obwohl* die zugrundeliegenden Informationen stets unvollständig und unsicher sind. Es stellt sich nun heraus, dass Zufallsverfahren für diese Aufgabe extrem hilfreich sind. Zufallsverfahren sind Algorithmen oder Prozeduren, die Zufallszahlen benutzen, um das Vorgehen in nicht-deterministischer Weise zu steuern. Solche Prozeduren sind in der Natur weit verbreitet, man denke nur an die große Vielfalt bei der Entstehung von Schneeflocken oder an die „somatische Hypermutation", mit der das Immunsystem von Tieren und Menschen schnell geeignete Antikörper gegen neue, bisher unbekannte Krankheitserreger finden kann. Auch in der menschlichen Gesellschaft werden häufig Zufallsverfahren verwendet. Ein klassisches Beispiel ist die Demokratie des antiken Athens, die sehr stark auf die Personalauswahl durch das Los (implementiert durch ein ausgeklügeltes Gerät

Abbildung: Nachbau eines Kleroterions (Quelle: Wikipedia, © CC0)

namens *Kleroterion*) und weniger auf Wahlen setzte (nur das erstere galt als „Volksherrschaft" im eigentlichen Sinne). Das heute verbreitetste Beispiel ist vermutlich das Konzept des „kontrolliert-randomisierten Experiments", das grundlegend für die Bewertung neuer Medikamente und medizinischer Behandlungsmethoden ist.

ALGORITHMEN UND ZUFALL

Am weitesten geht die Randomisierung in der Welt der Automaten, also in der künstlichen Intelligenz. Die modernsten Methoden des maschinellen Lernens, insbesondere die neuronalen Netze, sind ohne Zufallsverfahren undenkbar. Alle heute verwendeten Lernalgorithmen beruhen auf zufällig gewählten Trainingsdaten, die in zufälliger Reihenfolge verarbeitet werden, wobei die zu trainierenden Netze zufällig initialisiert und bei einigen Verfahren (z. B. *Drop-out*) auch noch in jedem Schritt zufällig modifiziert werden. Entsprechende deterministische Verfahren haben hinsichtlich der Qualität der Resultate keine Chance, diese Methoden zu schlagen.

Auch die modernen Ansätze zur Bestimmung des jeweiligen Grades der Unsicherheit im Rahmen der Bayes'schen Inferenz setzen in großem Maße auf Randomisierung, etwa durch die Verwendung von zufallsgesteuerten Simulationsprogrammen, Monte-Carlo-Sampling und Bayes'schen neuronalen Netzen.

Der unübersehbare Erfolg der Zufallsverfahren bei Automaten legt die Frage nahe, ob auch Menschen bzw. die menschliche Gesellschaft von solchen Verfahren profitieren können. Dies war eine Kernfrage der interdisziplinären Zusammenarbeit in unserem Team und in der Fellow-Gruppe insgesamt, weil jede vertretene Fachdisziplin jeweils einen besonderen, einzigartigen Blickwinkel auf die Frage beitragen konnte. Der Retreat zu Beginn der Fellowship war sehr hilfreich, um sich schnell kennenzulernen und diese Diskussionen zügig in Gang zu bringen. Ebenso förderlich war es, dass ein weiteres Team unter dem Titel „*Künstliche Intelligenz: Zwischen Wunderglaube und Wissenschaft*" einen durchaus kritischen und dadurch herausfordernden Ansatz zu eng verwandten Fragen vertreten hat.

INTERDISZIPLINÄRE PERSPEKTIVEN

Sehr aufschlussreich war für mich die von Stefan Trautmann vermittelte Begegnung mit Chengwei Liu, der bereits seit Jahren zufallsbasierte Verfahren in der Personalauswahl erforscht. Wenn mehrere Jobkandidat:innen hinsichtlich der Hauptkriterien gleich gut abschneiden, versucht man traditionell mit sekundären Kriterien eine rational begründete Rangfolge zu definieren. Es stellt sich aber heraus, dass die sekundären Kriterien nur geringe Vorhersagekraft für das wirkliche Potenzial der Kandidierenden haben und stattdessen eher die (häufig diskriminierenden) Vorurteile der Arbeitgeber:innen ausdrücken und prolongieren. Chengwei Liu konnte empirisch zeigen, dass eine Zufallsauswahl zwischen gleich geeigneten Kandidierenden zu besseren Entscheidungen führt, weil sie unnötige Sekundärkriterien und die damit verbundenen Verzerrungen vermeidet. Ähnliche Ideen werden bei der Auswahl von Artikeln für Publikationen und Konferenzen sowie von Forschungsanträgen verfolgt, die gleichfalls davon geprägt sind, dass viele Einreichungen (im Rahmen gewisser Fehlerschranken) ununterscheidbare Qualität haben. Allerdings sind diese Untersuchungen noch in der Pilotphase und haben sich nicht allgemein durchgesetzt.

Auch von einigen Politikwissenschaftler:innen wird die Frage von Zufallsverfahren in der Gesetzgebung intensiv diskutiert. Zufällig zusammengesetzte Bürgerräte ha-

ben in verschiedenen Ländern (unter anderem in den Niederlanden, Belgien, Kanada und Deutschland) fundierte Lösungsstrategien zu wichtigen gesellschaftlichen Fragen (darunter zur Bewältigung der Klimakrise) erarbeitet. Dabei wurde gezeigt, dass solche Gremien – bei geeigneter Expert:innenberatung – sehr leistungsfähig sein können und viel weniger als die üblichen Gremienformen unter Verzerrungen (Parteidisziplin, mangelnde Kenntnis der tatsächlichen „Volksmeinung", Group Think, Risikoaversion, etc.) leiden. Allerdings haben diese Experimente bisher kaum Konsequenzen, da ihre Empfehlungen in keiner Weise bindend sind. Es wäre deshalb lohnend, diese Ansätze – auch unter Berücksichtigung der Erkenntnisse des maschinellen Lernens – so weit zu verfeinern und theoretisch abzusichern, dass sie auch bei verbindlichen politischen Entscheidungen eingesetzt werden können.

Auf der anderen Seite hat Jan Schur als Jurist darauf hingewiesen, dass Zufallsverfahren in der Rechtsprechung enge Grenzen gesetzt sind, da Urteile rational begründbar und bei gleichem Sachverhalt auch gleich getroffen werden müssen. An diesen Blickwinkel knüpft die Hauptdiskussion mit Rebecca Müller und Andreas Voß in meinem eigenen Team an: Welche Rolle spielen Transparenz und Erklärbar-

keit für die Akzeptanz von Automaten durch Menschen? Arthur Clark's bekannte Aussage „Jede hinreichend fortschrittliche Technologie ist von Magie nicht zu unterscheiden" galt in gewisser Weise bereits im Mittelalter: Herrschende ließen sich sehr gern durch undurchschaubare Automaten göttergleich in Szene setzen. Allerdings bemerkt man auch, dass sich Chronisten schon damals häufig (und erfolgreich) um rationale Erklärungen des Gesehenen bemüht haben. Heute ist hier natürlich die Psychologie gefragt: Welche Art von Erklärungen über Automaten sind für Menschen besonders hilfreich? Wie lässt sich Zufall einsetzen, um die Unsicherheit von Entscheidungen anschaulich zu vermitteln (z. B. durch Monte-Carlo-Simulation verschiedener Szenarien)? Wann sind rationale Argumente (wie die oben erwähnten sekundären Kriterien bei der Personalauswahl) nur Scheinerklärungen? Wie kann man Zufallsverfahren so steuern (etwa durch eine gewichtete statt einer gleichverteilten Auswahl), dass sie nicht als irrational empfunden werden? Diese und ähnliche Fragen eröffnen ein weites Feld für die zukünftige Forschung, von der sowohl die künstliche Intelligenz als auch die menschliche Gesellschaft stark profitieren könnten.

AUSBLICK

Um diese Diskussionen zu vertiefen, bieten wir im kommenden Semester ein gemeinsames interdisziplinäres Brückenseminar „Mensch und Automat" für alle interessierten Studierenden der Universität an.[1] Die anvisierten Themen umfassen ein sehr breites Spektrum, beginnend mit Automaten in Geschichte und Kunst über künstliche Intelligenz in der Psychologie, nachvollziehbares und erklärbares maschinelles Lernen bis hin zu Zukunftsfragen bezüglich der Ethik „autonomer Agenten" und einer hypothetischen Intelligenzexplosion, die zu einer unwiderruflichen Überlegenheit der Automaten über den Menschen führen würde. Ich persönlich hoffe außerdem, dass das Seminar – unter anderem inspiriert durch den Rückblick auf das antike Athen – zu einer erneuerten gesellschaftlichen Diskussion über die Zukunft der Demokratie und der demokratischen Verfahren beiträgt. Angesichts der aktuellen Bedrohungen sowohl durch extreme politische Positionen als auch durch erstarkende autoritäre Staaten halte ich diesen Diskurs für dringend geboten.

[1] https://hci.iwr.uni-heidelberg.de/teaching/seminar_automata_2022.