»Erst die Frage, dann die Operationalisierung, dann die Daten – Nur so können Entscheidungen über Bildungsbemühungen getroffen werden«

Interview mit Katharina Anna Zweig

Abstract
Im Rahmen unseres Buchprojekts haben wir auch Kolleg:innen ganz unterschiedlicher Fachdisziplinen gebeten, uns zum Thema »Vermessung der Bildung« ihre Perspektiven näher zu bringen. Im Folgenden lesen Sie ein Interview mit Katharina Zweig, Informatikprofessorin an der RPTU Kaiserslautern-Landau. Die Literaturhinweise im Interviewtext wurden im Nachgang zur Vertiefung ergänzt.

Mandy Schiefner-Rohs (MSR): Liebe Katharina, wir kennen uns schon lange, aber für die Lesenden des Buches wäre es bestimmt interessant zu wissen, was du eigentlich machst.

Katharina Anna Zweig (KAZ): Grundsätzlich beschäftige ich mich mit der Frage, wie man etwas so messen kann, dass das Resultat als Grundlage für eine Entscheidung genutzt werden kann. In meiner Forschung geht es zudem häufig um Modellierungsentscheidungen. Ich habe mich beispielsweise jahrelang damit beschäftigt, wie man komplexe Netzwerke analysieren kann. Dabei ging es insbesondere um eine Gruppe von Maßen, die »Zentralitätsmaße« heißen und genau das tun sollen: die Wichtigkeit oder Zentralität von Knoten im Netzwerk zu bewerten. Davon gibt es ungefähr 60. Als Informatikerin habe ich mich lange gefragt, warum es davon so viele gibt und woher ich weiß, wann ich welches Maß verwenden soll. Und die Antwort, die Geistes- und Sozialwissenschaftler natürlich kennen, ist, dass die Maße jeweils unterschiedliche Konzepte oder Interpretationen von Wichtigkeit, von Zentralität repräsentieren. Aber als Informatikerin war mir das nicht so klar. In der Informatik wird dann oft gelehrt: Hier ist ein Netzwerk, hier sind 60 Maße – und das eine nimmst du, wenn du wissen willst, wer in einer Kommunikation zwischen anderen steht, das andere nimmst du, wenn du wissen willst, wie schnell jemand alle anderen im Netzwerk erreichen kann. Es ist also nur eine Verbalisierung der Formel, die dahintersteht, ohne zu sagen, dass das Zentralitätsmaß angepasst sein sollte auf

die Situation und auf das, was in dem Netzwerk geschieht. Dazu ein Beispiel: Eins der Zentralitätsmaße misst, wie sehr ein Knoten in dem Netzwerk als Vermittler zwischen anderen Knoten benötigt wird. Die Formel geht aber implizit davon aus, dass alle Paare von Knoten gleich oft miteinander »kommunizieren« wollen, selbst wenn sie im Netzwerk weit voneinander entfernt sind. Ein solches Maß sollte man daher nicht anwenden auf Netzwerke, die die menschliche Kommunikation darstellen und zig Millionen von Menschen in verschiedenen Ländern repräsentieren. Es reicht daher nicht, nur die Formel zu kennen, man muss auch den Kontext der Daten berücksichtigen, damit die Resultate einen Sinn ergeben. Aus meiner Sicht ist es eine Modellierungsentscheidung, wann man welches Maß zur Beantwortung einer Frage verwendet.
Mit solchen Fragen habe ich mich bis ca. 2018 hauptsächlich beschäftigt. In den letzten Jahren hat sich diese Frage darauf ausgeweitet, wann man mithilfe von Maschinen Entscheidungen treffen kann, insbesondere mit sogenannten Künstlichen Intelligenz (KI)-Systemen. Und auch das hat ganz viel mit Modellierungsentscheidungen zu tun. Denn maschinelles Lernen ist nicht von allein objektiv oder optimal, sondern auch hier entscheidet eine Person, welche Art von Daten in die Maschine hineingehen und mit welcher Methode »gelernt« wird. Oftmals gibt es dazu viele Parameter, die man dann noch einstellen kann. Mich interessiert, wie sich diese Entscheidungen auf die Interpretierbarkeit des Resultats auswirken.

MSR: Welches Schlagwort kommt Dir denn als Erstes in den Sinn, wenn Du an den Titel unseres Buches – Vermessung der Bildung – denkst?

KAZ: Wenn ich an den Titel des Buches »Vermessung der Bildung« denke, dann ist es die Frage nach der Operationalisierung, die mir dazu einfällt. Darunter verstehe ich die Entscheidung, wie man etwas messen will – das fällt ja nicht vom Himmel. Aus meiner Sicht ist die Idee, dass eine Formel mir helfen kann, eine komplexe Situation zu bewerten, ein Modell. Ein Modell definiere ich wie Michael Weisberg (2015): es hat eine Struktur und eine Interpretation. Er nennt die Interpretation des Modells das *Construal*. Es besteht bei ihm aus drei Teilen: Das ist einmal die *Zuweisung*, die angibt, welche Strukturelemente welche Elemente der Welt repräsentieren. Den zweiten Teil nennt er *Intended Scope,* den angedachten Anwendungsbereich: Wofür wurde das Modell primär designt? Und dann nennt er als dritten Teil noch *Fidelity*-Kriterien, also Glaubwürdigkeitskriterien oder Passungskriterien, die dir sagen sollen, ob das Modell auf eine andere Situation passt oder wie gut es darauf passt. Und die Anwendung eines Maßes erfüllt meiner Meinung nach diese Kriterien. Die Struktur selbst liegt in der Formel: Das von mir zuvor beschriebene Zentralitätsmaß geht davon aus, dass es auf diesem Netzwerk einen Netzwerkfluss gibt, der entlang kürzester Wege verläuft, und dass alle Paare von Knoten mit derselben Intensität an diesem Netzwerkfluss teilnehmen. Ein Beispiel dafür: Wenn das Netzwerk ein Kom-

munikationsnetzwerk zwischen Personen beschreibt, beispielsweise repräsentiert, wer wen anruft, dann wäre der Netzwerkfluss hier vielleicht die Menge an Informationen, die von Person A zu Person B fließen. Die kann natürlich auch indirekt fließen: A ruft C an, der B anruft – so fließt Information von A nach B über C. Das Zentralitätsmaß geht nun davon aus, dass alle Paare von Personen genau gleich viele Informationen austauschen wollen, unabhängig davon, wie viele intermediäre Personen zwischen ihnen stehen. Eine Person ist dann wichtig oder zentral, wenn sie für viele Paare auf vielen der kürzesten Kommunikationswege zwischen ihnen sitzt. Damit habe ich nun auch die Zuweisung beschrieben: Knoten im Netzwerk sind Personen, der Netzwerkfluss ist in diesem Fall Kommunikation per Telefon. Der angedachte Verwendungszweck ist für Situationen, die darauf passen, in der also auf diese Art kommuniziert wird. Nun ist es in der realen Welt natürlich nie genau so. Daher könnte ein Passungskriterium sein, wie oft Paare von Knoten wirklich miteinander kommunizieren und ob sie das auf den jeweils kürzesten Wegen tun. Wenn das nicht der Fall ist, sollte man das Maß nicht verwenden. All das gilt nun natürlich auch für die Vermessung der Bildung: Wenn wir etwas messen, wollen wir ja damit immer eine Dimension der Welt hinreichend erfassen, um dann Entscheidungen treffen zu können. Und gerade bei der Vermessung von Bildung geht es dann vielleicht um solche Entscheidungen wie: »Wo müssen mehr Ressourcen hin? Brauchen wir mehr Lehrkräfte? Brauchen wir eine andere Art von Bildung? Lernen unsere Kinder und unsere Studierenden das, was sie nachher brauchen?« Mit unseren Messungen wollen wir also meistens irgendeine Art von Entscheidung vorbereiten. Und die Frage ist dann, ob das Modell, das wir in die Formel legen, für diese Situation und diese Fragestellung angemessen ist. Und meistens passen unsere Formeln und Prozesse nicht zu diesem dritten Kriterium, nämlich den Passungskriterien. Wir machen uns oftmals zu wenig Gedanken darüber, woran jemand erkennen kann, ob die Situation so ist oder sich vielleicht so weit geändert hat, dass wir dieses Messinstrument gar nicht mehr verwenden sollten.

MSR: Ja. Ich versuche es mal so zu formulieren, ob ich es richtig verstanden habe, dass a) der Kontext dementsprechend relevant ist und wir bei der Vermessung von Bildung immer diesen Kontext auch mit in den Blick nehmen und dass es b) letztendlich immer auch eine konkrete Fragestellung braucht.

KAZ: Genau, denn ansonsten könnte man gar nicht feststellen, ob der angedachte Verwendungszweck des Messmodells, der Formel oder des Prozesses überhaupt getroffen ist. Und gerade im Bereich der KI-Systeme gibt es jede Menge Modellannahmen, die man eigentlich prüfen müsste. Das gilt natürlich auch dann, wenn wir damit Bildung vermessen.

MSR: Und eine solche Modellannahme bezogen auf Bildung ist ja auch die Vorstellung von Lernen, und Lehren und davon, wie wir eigentlich lernen.

KAZ: Genau das wird ja ganz selten explizit gemacht in Softwaresystemen. Aus meiner Sicht liegt das meistens an einer schlechten Kommunikation zwischen den Domänenexperten und -expertinnen und den Informatikern und Informatikerinnen. Wenn die Domänenexpertinnen und -experten keine klaren Vorstellungen äußern, wie die Software arbeiten soll, treffen die Entwicklungsteams Entscheidungen, ohne sich dessen bewusst zu sein, welche Auswirkungen das hat oder welche Annahmen hinter diesen Entscheidungen stecken.

MSR: Und inwieweit befasst Du dich selbst mit der Vermessung von Bildung?

KAZ: Ich habe mich zusammen mit Professor Jan Georg Schneider vom Campus Landau der RPTU mit der Frage beschäftigt, ob KI-Systeme verwendet werden können, um Essays zu benoten.[1] Warum sollte man das wollen? Naja, erstens sind Menschen natürlich fehlbar, das haben viele psychologische Studien schon gezeigt. Zweitens ist die Beurteilung eines Textes zeitintensiv. Insbesondere bei Massenprüfungen von Tausenden von Personen ist das ein Faktor. Dazu gehört die Beurteilung von Sprachessays, die von nicht Muttersprachlerinnen oder Muttersprachlern geschrieben werden, z.B. um den TOEFL-Test zu machen. Hier bekommen sehr viele Menschen gleichzeitig dieselben Aufgaben gestellt, die dann alle anhand desselben Maßstabes korrigiert werden müssen. Der Prozess an sich ist sehr aufwendig, wenn Menschen das machen. Das sind speziell ausgebildete Personen, die morgens z.B. erst einmal Testessays bewerten müssen. Nur dann, wenn sie diese gut genug bewerten, also konsistent genug mit all den anderen Gutachterinnen und Gutachtern, dürfen sie den Tag als Bewerter überhaupt anfangen. Zusätzlich bekommen sie den ganzen Tag über Testessays, wissen das aber nicht. Wenn sie dort zu stark abweichen, müssen sie ebenfalls aufhören. Es ist also wirklich ein sehr aufwendiger Prozess, der einer starken Qualitätskontrolle unterliegt. Und diesen Prozess wollte man jetzt mithilfe eines KI-Systems beschleunigen.
Ist das möglich? Dazu haben Jan Georg und ich uns mit einem Patent von 2002 auf ein solches Bewertungssystem beschäftigt, dem sogenannten E-Rater. Als Grundlage braucht diese Software ungefähr 250 bis 300 von Menschen korrigierte Essays, von denen »gelernt« wird, wie die Abgaben auf jeweils eine Fragestellung bewertet

1 Schneider, J.G. & Zweig, K.A. (2022). Ohne Sinn. Zu Anspruch und Wirklichkeit automatisierter Aufsatzbewertung. In S. Brommer, J. Spitzmüller & K.S. Roth (Hg.), Brückenschläge – Linguistik an den Schnittstellen. Narr Francke Attempto Verlag. https://elibrary.narr.digital/content/pdf/10.24053/9783823395188.pdf (abgerufen am 19.06.2023)

werden. Wenn die Maschine das gelernt hat, kann man ihr noch nicht bewertete Essays geben und sie sagt dann vorher, wie ein Mensch sie bewerten würde, basierend auf diesen 250–300 von Menschen korrigierten Essays. Das Lernen funktioniert so: Zuerst werden für alle menschlich benoteten Essays ziemlich triviale Dinge gemessen, also so etwas wie: »Wie viele Modalverben, also so etwas wie *can* und *will*, wurden verwendet?«. Davon gibt es eine Handvoll Fragen, je nach der genauen Textart, die in der Prüfung verlangt wird. Interessanterweise wird z.B. bei Pro-Contra-Argumentationen im letzten Paragraphen gezählt, wie viele Modalverben im Konjunktiv gesetzt sind und die Anzahl von Nebensätzen pro Satz bestimmt. Das sind alles einzelne Zahlen. Dann generiert man noch zwei weitere Zahlen, die schon direkt Noten sind. Dazu komme ich gleich. Am Ende hat man eine Handvoll Zahlen, die alle das Ergebnis von sehr einfachen Rechenverfahren und Vergleichen sind. Der KI-Aspekt ist jetzt, dass die Maschine versucht zu lernen, wie man diese Zahlen miteinander gewichten muss, um die von den Menschen vergebene Note vorhersagen zu können. So, that's it. Mehr macht man da nicht.

Die beiden Zahlen, die direkt Noten sind, sind besonders interessant: Die eine Zahl kommt daher, dass man sich die Worthäufigkeitsnutzung anschaut von einem zu bewertenden Essay und diese vergleicht mit der Wortnutzungshäufigkeit von allen Essays, die von den Menschen mit einer 1 bewertet wurden, mit einer 2 mit einer 3 und so weiter. Und da, wo die Wortnutzungshäufigkeit am ähnlichsten ist, ist die Note dann ebenfalls eine Zahl, die in die gelernte, gewichtete Formeln eingeht. Dann gibt es noch eine Zahl, bei der man die Worthäufigkeitsnutzung für jedes einzelne Argument in dem Essay berechnet, mit den Worthäufigkeiten in den Notenklassen vergleicht, die am besten passende Notenklasse auswählt und über alle Argumente den Durchschnitt bildet. Oft ist es dann so, dass sehr gut bewertete Essays seltenere Wörter verwenden – ebenfalls wieder ein Hinweis auf ein höheres Sprachniveau.

Ich habe so etwas Ähnliches auch einmal für eine von meinen eigenen Klausuren gemacht. Ich habe 16 Klausuren gehabt, die ersten acht korrigiert, und dann basierend auf einer einzigen Eigenschaft der Klausuren eine KI trainiert. Mit dieser KI habe ich dann eine Vorhersage gemacht, basierend auf dieser einen einzigen Eigenschaft, welche Note die anderen acht verbliebenen Klausuren bekommen werden. Und tatsächlich habe ich im Durchschnitt nur um 0,3 daneben gelegen. Aber was war diese eine einzige Eigenschaft der Klausur? Es war die Anzahl der eingereichten Seiten. Und siehe da, es war ein hervorragender Prädiktor für die Note. Das geht bei meinen Studierenden ganz gut, denn sie schreiben eben nur dann viele Seiten, wenn sie vorher entsprechend gut recherchiert haben. Daher korreliert die Seitenanzahl mit dem Rechercheaufwand und ist – anscheinend – ein ganz guter Prädiktor für die nachher von mir aufgrund der inhaltlichen Qualität der Arbeit vergebenen Note. Aber wenn ich ihnen im Vorhinein gesagt hätte, dass ich ihnen die Note anhand der Anzahl der eingereichten Seite gebe, dann wäre die-

se Korrelation zwischen der Seitenanzahl und der inhaltlichen Qualität natürlich verlorengegangen.
Beide Beispiele zeigen, dass man sehr einfache Dinge zählen kann und dass man damit manchmal tatsächlich ganz gut vorhersagen kann, welche Note ein Mensch gibt. Das gilt aber nur, solange sie mit der Note *korrelieren*, so wie es hier mit der Anzahl der eingereichten Seiten der Fall war. Eine sehr einfache Eigenschaft eines Textes kann also eine ziemlich treffsichere Notenvorhersage ermöglichen, auch wenn das Gemessene die inhaltliche Qualität selbst gar nicht bewertet. So ähnlich ist es auch beim E-Rater: Die Anzahl von Nebensätzen pro Hauptsatz ist mit der Note stark korreliert, da es ein fortgeschrittenes Sprachkonstrukt ist, das man sich als Nichtmuttersprachler erst dann traut, wenn man ein gewisses Niveau hat. Genau wie die Verwendung des Konjunktivs von Modalverben. Das sind also alles Anzeichen für ein höheres Sprachniveau, solange es die Prüflinge nicht allein deswegen tun, weil sie wissen, dass die Maschine darauf positiv reagieren wird. Das kann also nur funktionieren, solange die Prüflinge davon ausgehen, dass ein Mensch den Text korrigiert und dieser Mensch den Text auch inhaltlich verstehen muss. Wenn man aber davon ausgeht, dass nur die Maschine den Text liest, dann wäre es eben ausreichend, wenn man einfach viele Nebensätze, viele Dinge im Konjunktiv schreibt und ganz viele seltene Wörter verwendet. Insofern kann eine maschinelle Vorhersage einer Note die eigentliche inhaltliche Bewertung nicht ersetzen, aber vielleicht unterstützen. Das ist mein Berührungspunkt mit der Vermessung von Bildung.

MSR: Optimales Lernen, Optimierung des Lernens oder die optimale Gestaltung von Bildungsorganisationen – worum geht es deines Erachtens, wenn von der Vermessung von Bildung gesprochen wird?

KAZ: In meiner Community wird eher hervorgehoben, dass es damit eine Personalisierung des Lernens geben könnte, und Personalisierung wird dann als Möglichkeit zur Optimierung gesehen. Damit wird dann begründet, warum wir solche KI-Systeme bräuchten. Stellen wir uns vor, dass es wirklich die Möglichkeit gäbe, mithilfe eines Systems jederzeit gut begründete Rückmeldungen zum Lernfortschritt zu bekommen. Das wäre sicherlich etwas, was man sich wünscht. Ich stelle aber in Frage, ob das mit den heutigen KI-Systemen geht. Diese Frage war noch ziemlich einfach zu beantworten vor November 2022, weil die Systeme, die man da sehen konnte, eher so waren wie der E-Rater. Da wurde also etwas sehr Grobes gemessen, das gut korreliert mit dem, was man eigentlich messen will. Und deswegen hat das einigermaßen funktioniert, dass die Systeme eine Rückmeldung gaben, aber begründen konnten sie diese Rückmeldung nicht. Die Begründung ist aber das, was wir als Menschen brauchen, um zu verstehen, ob eine Rückmeldung an uns als Lernende hilfreich und sinnvoll ist. Jan Georg und ich haben bei unserem For-

schungsprojekt versucht herauszufinden, inwieweit eine Benotung durch die Maschine im Sinne der Sprechakttheorie von Austin als gelungen betrachtet werden kann. Ein Sprechakt ist ganz allgemein ein Satz, der nicht richtig oder falsch ist, sondern der mehr oder weniger gut gelingen kann und der etwas tut. Bekannte Beispiele sind Schiffstaufen oder Vermählungen: Durch Sätze, die in einem bestimmten Kontext von bestimmten Personen gesprochen werden, bekommt ein Schiff einen Namen und zwei Menschen gelten danach als Eheleute. Eine Benotung ist auch so ein Sprechakt: In dem Moment, in dem ich zu einem Prüfling sage: »Das war eine sehr gute Prüfung«, weise ich ihm die Note zu. Zu einem gelungenen Sprechakt gehört, dass der Sprecher oder die Sprecherin das Gesagte auch ernst meint. D.h., wenn ich eine Note gebe, dann kann ich diese Note inhaltlich begründen; ich stehe zu der von mir vergebenen Note und meine Reputation als faire Prüferin ist davon abhängig. All das kann die Maschine nicht, sondern sie gibt eine Note, die zwar genauso aussieht wie eine, die von einem Menschen vergeben wird, die aber weder das Resultat desselben Bewertungsprozesses ist, noch inhaltlich begründet werden kann. Denn sie beruht ja gar nicht auf der inhaltlichen Qualität, sondern auf charakterisierenden Eigenschaften wie der Anzahl der Nebensätze pro Satz.
Jetzt haben wir ChatGPT und die Welt ändert sich ein bisschen. Denn tatsächlich kann man z.B. ChatGPT bitten, eine Begründung zu schreiben. Man kann ihm also einen Essay geben und dann sagen: »Jetzt bewerte mir doch bitte mal diesen Essay und begründe die Note.« Und dann schreibt er einen Text und dieser liest sich auch wie eine Begründung, aber wenn man genau hinschaut, hat das, was bemängelt wird oder das, was gelobt wird, mit dem spezifischen Text nicht so viel zu tun. Das kann man schnell daran sehen, dass man ChatGPT dazu bringen kann, für ein und denselben Text eine sehr gute Note zu vergeben, eine mittlere und eine sehr schlechte Note. Jedes Mal wird sich die Begründung an sich gut lesen, aber die bemängelten Fehler werden im Text nicht zu finden sein. Und d.h., es fehlt uns also immer noch eine Maschine, die nicht nur etwas generieren kann, das wie eine Begründung aussieht, sondern auch wirklich eine ist, die sich auf den vorliegenden Text bezieht. Und deswegen hat sich die Welt mit Erscheinen der generativen KI-Systeme auf der einen Seite radikal verändert, weil wir zum ersten Mal maschinengenerierte Texte sehen, die verschiedene Textformen kennen und diese sehr gut reproduzieren können. An dem grundsätzlichen Unvermögen der Maschinen, Texte zu bewerten und zu beurteilen, hat sich allerdings nichts geändert. Sie verstehen das Gesagte nicht.

MSR: Nicht oder noch nicht. Also meinst du, ist es überhaupt möglich?

KAZ: Man sieht mit den großen Sprachmodellen jetzt sehr deutlich, dass der Begriff Verstehen mindestens zwei Aspekte beinhaltet: Das Wissen, wann und in welchem Kontext ein Wort sinnvoll gebraucht werden kann, und das Wissen, welches physi-

sche Objekte oder welche erfahrbare Situation das Wort bezeichnet. Bei Menschen ist das meistens miteinander verbunden, aber manchmal haben auch wir nur den ersten Aspekt eines Wortes »verstanden«. Im Englischen geht es mir z.B. manchmal so, dass ich plötzlich ein Wort benutzen will, von dem ich nicht aktiv weiß, was die Bedeutung ist, aber von dem ich weiß, dass man es in diesem Kontext verwenden kann. Und wenn ich es dann nachgucke, bin ich manchmal überrascht, dass es wirklich gut passt. Es war ein rein passives Wissen, das aus dem Lesen oder Hören des Wortes in bestimmten Kontexten kommt.
Dieses letztere »Wissen« haben die großen Sprachmodelle jetzt – in einem rein statistischen Sinne. Solange wir bei diesen Arten von Technologien bleiben, also dem reinen maschinellen Lernen, werden wir aus meiner Sicht über dieses Stadium des »Wissens« oder »Verstehens« nicht hinauskommen. Gary Marcus und Ernest Davies sagen in ihrem Buch »Rebooting AI« aus dem Jahr 2019, dass wir dafür eine neue Art der Technologie benötigen, die sowohl Weltwissen als auch maschinelles Lernen miteinander vereint. Der Philosoph Brian Cantwell Smith kommt 2019 zu einem ähnlichen Ergebnis und schlägt vor, Wörter wie »Wissen« oder »Verstehen« – wenn sie Maschinen bezeichnen – wie folgend zu kennzeichnen: ⌈Wissen⌉, ⌈Verstehen⌉. Aber gerade beim Zweiteren rät er im Allgemeinen davon ab, es im Zusammenhang mit Computern zu verwenden.
Daher ist meine Antwort: Mit der heutigen Technologie eher nicht, aber mit anderen Technologien ist es natürlich denkbar. So argumentiere ich auch in meinem neuen Buch (Zweig, 2023).

MSR: Und das führt uns gleichzeitig zur ersten Frage zurück: Der Kontext wird wieder relevanter.

KAZ: Ja, und genau den Teil hat die Maschine jetzt besser ⌈verstanden⌉. Wie schon gesagt, wird der Kontext dabei nicht in dem Sinne umfassend erfasst und verstanden, wie es Menschen tun. Aber er wird mit der heutigen Technologie so gut berücksichtigt, dass es fast so wirkt, als würde der Kontext verstanden. Ich denke, dass dies eine ganz gute Formulierung ist. Ich glaube daher, dass man auch mit den Sprachmodellen, die wir heute schon haben, ein hilfreiches Feedback auf der grammatischen Ebene bekommen kann. Mein Italienisch ist z.B. sehr rudimentär und da würde mir ein Italienisch-Chatbot mit viel Geduld und sehr guter Grammatik sehr viel bringen. Da wäre es dann auch egal, ob die Maschine jetzt halluziniert, also ob alles Gesagte jetzt immer inhaltlich korrekt ist. Ich will mit der Maschine ja nicht über italienische Politik oder die Geschichte Italiens parlieren, sondern mein Alltagsitalienisch wieder auffrischen. Und allein dadurch, dass jemand auf der anderen Seite mit unendlich viel Geduld sitzt, der mir zu 95 % mit sinnvollen Sätzen antwortet, lerne ich weiter. Ich verstehe heute einiges, wenn ich es lese und höre, vielleicht 30–40 Prozent, und könnte mit einem geeigneten Chatbot vielleicht auf 80 bis

90 Prozent Leseverständnis kommen. Mein Ungarisch ist dagegen so schlecht, dass ich noch nicht von einem solchen Chatbot profitieren könnte. Ich glaube, dass wir beim Sprachenlernen insbesondere für den Einstieg und den Feinschliff menschliche Anleitung benötigen. Aber der Teil dazwischen, der könnte jetzt meiner Meinung nach durch solche KI-Systeme stark beschleunigt und viel effizienter gemacht werden. Aber natürlich brauchen wir dafür Studien, um zu sehen, ob das wirklich so ist.

MSR: Und damit kann ich wieder Ressourcen einsparen, weil ich die Menschen in diesem Prozess dann nicht einsetze, sondern nur noch vorne und hinten an der Stelle.

KAZ: Man kann das natürlich so formulieren, aber das klingt so negativ. Man könnte auch sagen, dass es damit viel mehr Menschen ermöglicht wird, andere Sprachen zu lernen, weil es günstiger wird, Sprachen auf einem solchen Niveau zu lernen. Welcher Effekt überwiegt, wird sich dann zeigen.

MSR: Ein Stichwort, dass bisher noch gar nicht gefallen ist, ist das der Datafizierung: Inwieweit trifft dieses Phänomen heute den Kern der bereits länger andauernden Diskussion um die Vermessung von Bildung?

KAZ: Wieso habt ihr euch für den Begriff Datafizierung und nicht Digitalisierung entschieden? Was ist für euch der Unterschied?

MSR: Digitalisierung und Datafizierung hängen natürlich zusammen. Uns interessiert aber vor allem der letzte Aspekt, die Rolle einzelner Daten und Datenpunkte, die auch miteinander verknüpft werden können. Denn Daten beeinflussen ja aktuell viele Entscheidungs- und Meinungsbildungsprozesse – individuell, organisational und gesellschaftlich. Ganz stark aber betont zumindest aus unserer Perspektive der Datafizierungsbegriff, dass Daten beispielsweise erzeugt werden durch Digitalisierung, die vorher vielleicht nicht da waren, also Metadaten, Kontextdaten, lokal basierte Daten. Ich weiß jetzt, wann ein:e Student:in in der Mensa ist oder in der Bibliothek ist und solche Dinge. Und diese Daten können ja auch für Lehr-Lernprozesse genutzt werden. Daher war das unser Fokus im Projekt, und nicht so sehr die Digitalisierung an sich, wobei sich beide Perspektiven natürlich beeinflussen.

KAZ: Gut, aber für mich hat Datafizierung eine negative Konnotation, nämlich, dass über das Maß hinaus Daten erhoben werden. Als Informatikerin ist die Sichtweise eher so, dass durch die Digitalisierung plötzlich alles zum Datum wird, weil ich es erst dadurch verarbeiten kann. Und deswegen finde ich den Begriff der Digitalisierung hier neutraler. Alles, was digital ist, ist damit ein Datum und ist verar-

beitbar. Datafizierung bezeichnet für mich, dass alles, was Verhalten ist, auch zum Datum gemacht werden muss. Und das ist natürlich nicht der Fall. Wir müssen uns gut überlegen, was wir messen wollen. Und das muss Sinn ergeben. Damit sind wir wieder bei der den Fragen: Wie messen wir was und warum tun wir das? Und dazu muss natürlich immer die Frage zuerst geklärt werden: Was ist denn überhaupt der Schmerz? Was soll denn überhaupt geheilt werden? Welche Entscheidung wollen wir treffen? Und ich glaube, das ist so ziemlich genau das Gegenteil von dem, wie ich den Begriff Datafizierung verstehe. Denn Datafizierung stellt das Datum oder die Herstellung oder die Messung von Daten in den Vordergrund. Während ich mich immer frage: Was wollen wir denn eigentlich tun? Was ist das Ergebnis, was wir nachher haben wollen? Und insofern, wenn man Datafizierung als den Kern der bereits länger andauernden Diskussion um die Vermessung von Bildung sehen würde, dann wäre es aus meiner Sicht genau der falsche Ansatzpunkt. Denn wir sollten uns fragen, was Bildung heißt. Wir sollten uns fragen, warum wir Bildung nicht so umsetzen, wie wir es eigentlich für richtig halten, wenn wir unendliche zeitliche und finanzielle Ressourcen hätten. D.h., man braucht die Nebenbedingungen, unter denen Bildung passiert. Und unter diesen Nebenbedingungen kann man sich dann die Aufgabe stellen, dass man die individuellen Lernziele optimieren möchte, dass man möglichst effizient sein möchte oder sonst etwas. Und dann sollte man sich fragen, welche Daten man dafür braucht und wie man das messen will.

MSR: Du würdest auch nicht sagen, dass Datafizierung »passiert«, sondern Datafizierung ist aus deiner Sicht etwas Aktives: Jemand sammelt aktiv Daten und über das Maß hinaus. Das würde ja fast mit Sammelwut in Verbindung gebracht werden können. Denn eigentlich könnte man ja auch sagen, dass allein durch die Nutzung digitaler Medien Daten anfallen. Wo bin ich, wie schnell laufe ich, wie lang sind meine Schritte, bin ich im Lot und vieles mehr, was mit dem Smartphone mittlerweile gemessen wird bzw. werden kann. Für dich wäre es erst dann Datafizierung, wenn diese Daten aktiv verarbeitet werden, die Erzeugung würdest du noch nicht dazu zählen?

KAZ: Nein, das trifft es nicht ganz. Daten fallen nie irgendwo einfach so an. Immer hat sich irgendeine Person entschieden, dass das Gemessene wichtig genug ist, um es zu speichern. Daher ist es immer etwas Aktives, es kann nichts Passives sein. Und dann gibt es eine zweite Person, die die Daten als Grundlage verwendet, um daraus beispielsweise Bildungsaussagen zu machen. Aus meiner Sicht sollte aber die Reihenfolge anders sein. Die Datafizierung, also das reine Aufnehmen von Daten sollte eigentlich erst der dritte Schritt sein. Der erste Schritt sollte immer die Frage sein: Was wollen wir eigentlich wissen, welche Entscheidungsgrundlage benötigen wir? Der zweite Schritt besteht in der Beantwortung der Frage: Wie wollen wir das messen? Dann sollte man im dritten Schritt entscheiden, welche Daten man braucht.

Und dann kann man das Glück haben, dass diese Daten schon da sind, weil jemand diese Informationen schon gespeichert hat. Oder man stellt fest, dass die eigentlich gewünschten Informationen nicht vorhanden sind, aber ein Proxy dafür zu finden ist. Ein Proxydatum ist eine Information, die man als Approximation für das eigentlich Gewünschte verwenden kann.
Aber die zufällig vorhandenen Daten sollten eigentlich nie der Startpunkt sein, um grundlegende Entscheidungen über Bildungspolitik zu fällen. In meinem Buch Network Analysis Literacy (Zweig, 2016) unterscheide ich daher bei der Datenanalyse zwischen der explorativen und der zielgerichteten. Denn natürlich kann man mit Daten, die einfach irgendwo herumliegen, explorativ arbeiten: Steckt da Information drin? Das ist natürlich das, was wir als Wissenschaftler und Wissenschaftlerinnen den ganzen Tag tun, dass wir erst einmal Hypothesen bilden. Aber wenn man wirklich eine Entscheidungsgrundlage für Bildungspolitik oder für andere Bildungsfragen bilden möchte, dann darf man diesen Weg nicht gehen, sondern dann muss man eben zuerst die Frage stellen, dann die Operationalisierung festlegen und erst dann die Daten dafür sammeln.

MSR: ›All is data‹ – so heißt ja unser Forschungsprojekt in Anlehnung an Barney Glasers Plädoyer der qualitativen Sozialforschung. Stimmst du eigentlich zu?

KAZ: Das kommt jetzt drauf an, darüber kann man sich natürlich streiten. Oder man kann den Geist dessen, was da gesagt wird, bewerten. Am Ende ist natürlich alles, was unsere Sinne erreicht, in irgendeiner Form etwas, das wir als Grundlage für eine Theorie über die Welt verwenden können. Und wenn man sagt, dass alles, was als Input für eine Weltentheorie gilt oder benutzt werden kann, Daten sind, dann kann man ihm nur zustimmen. Das schließt aber eben das nicht mit ein, was nicht über Sinne direkt erfassbar ist. Und da wäre dann die Frage, inwiefern beispielsweise inneres Erleben auch sinnvoll als »Daten« bezeichnet werden kann. Ergibt ein solcher Satz wirklich Sinn? Sicherlich kann man also das, was extern auf uns einprasselt und irgendwie durch unsere Aufmerksamkeitsfilter kommt, als Daten bezeichnen. Ein anderer Satz, der mich mehr beschäftigt hat, ist: »Raw Data ist ein Oxymoron«, angelehnt an Geoffrey Bowker. Und der gefällt mir sehr gut. Das finde ich eigentlich viel interessanter, dass in dem Moment, wo wir unseren Blick auf etwas werfen, die »Rohheit« der Daten schon verloren gegangen ist. »Data is always cooked«, sie sind immer aus einem bestimmten Blickwinkel auf die Situation schon ausgewählt worden. Es gibt also keine unausweichliche, rein objektive Betrachtungsweise einer Situation mithilfe von Daten. Nur weil etwas ein Datum ist, ist es nicht »die Wahrheit« über diese Situation, sondern es ist das Ergebnis eines Messprozesses, auf den wir uns festgelegt haben. Er ist idealerweise als Messprozess intersubjektiv. D.h., ob ich ihn durchführe oder ob du ihn durchführst oder noch jemand anderes ihn durchführt: Wir können uns auf das Ergebnis einigen, und wenn das so ist, dann nennen

wir das Resultat einen Fakt. Aber wir wissen beide, dass die meisten Messprozesse nicht so fantastisch intersubjektiv sind, insbesondere dann, wenn Werturteile wie beispielsweise Noten mit dabei sind. Und je weniger intersubjektiv sie sind, desto weniger Faktenhaftiges liegt an diesen Daten. Insofern müssen wir bei allen Interpretationen von Daten immer mitberücksichtigen, wer aus welchen Gründen vorgeschlagen hat, dass genau die hinter den Daten stehenden Messprozesse geeignet seien und wie intersubjektiv diese Messprozesse sind. Insbesondere, wenn wir darauf basierende, gravierende Entscheidungen in der Bildungspolitik fällen.

MSR: Vielen Dank für deine Zeit und die interessanten Antworten!

Literatur

Bowker, G. C. (2008). Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care. In Ders. *Memory Practices in the Sciences*, MIT Press.

Marcus, G. & Davies, E. (2019). *Rebooting AI – Building Artificial Intelligence we can Trust.* Pantheon Books.

Schneider, J. G. & Zweig, K. A. (2022). Ohne Sinn. Zu Anspruch und Wirklichkeit automatisierter Aufsatzbewertung. In S. Brommer, J. Spitzmüller & K. S. Roth (Hg.), *Brückenschläge – Linguistik an den* Schnittstellen. Narr Francke Attempto Verlag. https://elibrary.narr.digital/content/pdf/10.24053/9783823395188.pdf (abgerufen am 19.06.2023)

Smith, B. C. (2019). *The Promise of Artificial Intelligence – Reckoning and Judgment*, MIT Press.

Weisberg, M. (2015). *Simulation and Similarity – using models to understand the world*. Oxford University Press.

Zweig, K. A. (2016). *Network Analysis Literacy.* Springer.

Zweig, K. A. (2023). *Die KI war's. Von absurd bis tödlich: Die Tücken der künstlichen Intelligenz.* Heyne.