kunft: Wie wird es uns ergehen, als zukünftige Cyborgs? Dieser Buchteil wird durch weitere Forschungsfragen im Bereich der Sozioinformatik ergänzt.

Nicht nur Technik wirkt sich unmittelbar auf Gesellschaft aus – auch Sprache hat diesen Effekt. Das lässt sich auch in Studien nachweisen [Stahlberg and Sczesny, 2001]. Daher haben wir uns in diesem Buch bewusst darum bemüht, Männer* und Frauen* gleichermaßen anzusprechen.

Alle Themen im Buch betreffen regelmäßig die verschiedensten Wissenschaftsfelder: Von der Soziologe über die Rechts-, Politik- und Wirtschaftswissenschaften bis hin zur Psychologie und Philosophie. Wir konnten diese nur streifen und hoffen daher, dass die vorgelegten Methoden und Überlegungen in allen beteiligten Disziplinen auf fruchtbaren Boden stoßen. Die von uns beschriebenen Beispiele zeigen, wie oft Entwickler:innen bei der Gestaltung von Hard- und Software Entscheidungen treffen, die besser geworden wären, wenn sie gemeinsam mit Vertretern und Vertreterinnen der oben genannten Wissenschaftsfelder und mit Bürgerinnen und Bürgern getroffen worden wären. Zu einer solchen gemeinsamen Gestaltung der digitalen Transformation möchten wir mit dem vorliegenden Buch beitragen.

Kaiserslautern, Stuttgart & Berlin, 10.03.2021

Katharina A. Zweig, Tobias D. Krafft, Anita Klingel und Enno Park

1 Warum Sozioinformatik – und warum jetzt?

Dass die Digitalisierung immer stärker in unsere Leben eingreift, ist schon kaum mehr als eine Platitüde. Wenn es aber um die Frage geht, welche Auswirkungen bei dem Einsatz eines neuen Software- oder Hardwareprodukts zu erwarten sind, fällt es den meisten doch schwer, eine detaillierte Technikfolgenabschätzung abzugeben, die klar macht, wo und wie sehr die Digitalisierung in unseren Leben Veränderungen auslösen wird. Die letzten Jahre haben stattdessen gezeigt, dass die Nutzer:innen, Regierungen, die Gesellschaft als Ganzes immer wieder überrascht werden von positiven wie negativen Nebenwirkungen von, beispielsweise, neuartigen digitalen Geschäftsmodellen, den Möglichkeiten von 3D-Druck z. B. für die Herstellung günstiger Prothesen, aber auch illegaler Waffen oder den neuen Manipulationsmöglichkeiten in der virtuellen Welt. In diesem Buch geht es um die Frage danach, ob und wie eine solche Technikfolgenabschätzung strukturiert durchgeführt werden kann. Auf den ersten Blick scheint es kaum möglich zu sein, bei solch breiten Anwendungsszenarien der Digitalisierung überhaupt einen Blick in die Glaskugel werfen zu können.Wir präsentieren hier eine Methode, die zeigt, dass eine strukturierte Analyse der Anreize für menschliches Verhalten, die durch den Einsatz einer Software verändert werden, es ermöglicht, einige Reaktionen von Menschen auf diesen Einsatz zu modellieren, zu analysieren und teilweise auch vorherzusagen. Und dazu betrachten wir als erstes Beispiel Software, die insbesondere in den USA und anderen englischsprachigen Ländern verwendet werden, um Essays von Schülerinnen und Schülern oder Studierenden zu bewerten.

1.1 Automatische Essaybewertung – die Zukunft objektiver und effizienter Benotung von Prüfungsleistungen?

Insbesondere im amerikanischen Bildungssystem spielen Essays als Prüfungsleistung eine große Rolle. Kaum ein Aufnahmetest an einer Hochschule kommt ohne sie aus, und auch in Studiengängen an deutschsprachigen Universitäten sind Aufsätze und Referate gängige Prüfungsformate. Das Problem daran: Solche Texte inhaltlich zu bewerten ist arbeitsintensiv und letzten Endes gibt es stets eine subjektive Komponente: Anders als beispielsweise Multiple-Choice-Tests gibt es bei Essays und akademischen Texten neben den faktischen Anteilen immer Komponenten, bei denen sich die Geschmäcker unterscheiden. Entspre-

chend müssen Anbieter solcher Prüfungen viel Zeit und Geld in qualifizierte Kräfte investieren, die Tausende von Essays korrigieren und bewerten. Insbesondere bei Tests wie dem *Graduate Record Examination* (GRE) oder dem *Test of English as a Foreign Language* (TOEFL), die in den USA häufig auch über Universitätszulassungen entscheiden, müssen die Noten zudem so objektiv und vergleichbar wie möglich vergeben werden, um Klagen zu vermeiden.

Ohne Frage hätte also eine automatische Bewertung von Essays, so sie denn möglich wäre, mindestens zwei Vorteile:

1. **Effizienz**: Sie wäre schneller und auf Dauer deutlich günstiger als menschliche Bewertungen.
2. **Fairness**: Sie würde alle Texte auf exakt dieselbe Art und Weise behandeln, wäre somit konsistent und würde daher – so die gängige Schlussfolgerung – niemanden bevorzugen.

Basierend auf dieser Argumentation ist es nachvollziehbar, dass der Bildungssektor seit Jahrzehnten darauf hofft, dass Algorithmen diese Arbeit eines Tages übernehmen können. Seit mehr als zwanzig Jahren werden immer wieder Patente angemeldet und wissenschaftliche Studien veröffentlicht, die genau dies versprechen (so bspw. [Burstein et al., 1998, Cahill et al., 2018]). Zusammengefasst wird die Diskussion um Potenzial und Grenzen solcher Systeme unter dem Schlagwort „Automated Essay Scoring“. Sie alle nutzen einen Bewertungsalgorithmus, der den digitalisierten Text auf Basis vorher festgelegter Kriterien überprüft und daraus eine Bewertung ableitet.

Übungsaufgabe 1:

Diskutieren Sie, welche Auswirkungen es haben könnte, wenn überall dort, wo Essays verlangt werden, diese von Computern bewertet würden. Welche Chancen und Vorteile würden sich ergeben? Sehen Sie Nachteile oder Risiken? Was könnte auf lange Sicht geschehen?

Um diese Fragen beantworten zu können, ist es sinnvoll, sich den Gesamtprozess anzusehen, in dem automatische Essaybewertungssysteme eingesetzt werden: Zunächst erhalten Lernende die Aufgabe, einen Essay zu einem vorgegebenen Thema mit einer vorgegebenen Wortzahl zu schreiben und einzureichen. Insbesondere in Tests wie den beiden oben genannten oder bei den Aufnahmeprozessen der großen Elite-Universitäten sind dabei die Aufgaben stark standardisiert und die erwartete Textstruktur klar vorgegeben. Bei einer Aufgabe, in der ein kontroverses Argument diskutiert werden sollen, würde eine Einleitung erwartet werden, Argumente für Pro und Contra, eine dialektische Synthese der Argumente und eine zusammenfassende Schlussfolgerung. Zudem sind meistens klare Grenzen für die minimale und maximale Wortzahl angegeben. Wird der Essay von menschlichen Gutachter:innen bewertet, dann sind dies in den meisten Fällen nicht die Lehrpersonen, sondern extra für die Bewertung geschultes Fachpersonal, das die Kandidat:innen nicht kennt: Die Gutachter:innen werden insbesondere angewiesen, neben der sprachlichen Korrektheit die Argumentationsstruktur und die Form des Textes zu bewerten. Fakten werden auf ihre inhaltliche Korrektheit dagegen nicht überprüft [Anson and Perelman, 2017]. Die starke Strukturierung der Aufgabe und des Essays kommt dabei auch der Bewertung durch eine

Software entgegen. Das Problem bei den automatische Essaybewertungssystemen ist aber, dass sie an die Grenzen dessen stoßen, wie Maschinen Informationen verarbeiten können. Wie alle Informatiker:innen wissen, müssen Informationen hierzu digitalisiert werden, also in Zahlenform erfassbar sein. Einige Aspekte eines Essays lassen sich dabei leichter digitalisieren als andere: So kann die Textlänge leicht als Anzahl der Zeichen oder Wörter bestimmt werden. Ob der Text allerdings auch überzeugend ist, ja sogar, ob er einfach nur faktisch korrekt ist, kann ein Computer nicht bewerten, auch nicht mit Methoden der sogenannten künstlichen Intelligenz.

Aber vielleicht muss der Computer das auch gar nicht können, solange er nur möglichst viele Texte mit möglichst genau derselben Note bewertet, wie es menschliche Gutachter:innen täten. Stattdessen nutzen daher automatische Essaybewertungssysteme leicht quantifizierbare Eigenschaften des Textes, sogenannte „features“, deren Werte mit den Bewertungen menschlicher Prüfer:innen korrelieren. Dazu gehört beispielsweise die Anzahl ungewöhnlicher Substantive oder die Länge einzelner Sätze. Beide korrelieren positiv mit der Höhe der von Menschen vergebenen Noten. Daneben suchen die Bewertungssysteme auch nach Stichwörtern, die anzeigen, dass hier Argumente gegeneinander abgewogen werden, wie „similarly“, „additionally“, oder „in contrast to this“. Ob diese Abwägung inhaltlich konsistent ist, kann aber nicht automatisch bewertet werden. Tatsächlich ist eine solch mechanische Bewertung immer noch der Stand der Dinge. Das folgende Zitat stammt aus einem Review von Januar 2020:

> *Another major characteristic of (automatic essay grading, AES) algorithms is that length matters. [...] Mark D. Shermis, a strong advocate of AES, reported that he has run the Gettysburg Address, what Gary Wills (1992) has called "the words that remade America", through several early AES machines, and the 271 word document received only 2s and 3s (Bloom, 2012).*
> *The focus on length is a central element of e-rater's algorithm for calculating a holistic score. Development and organization are calculated simply by counting the number of discourse elements (the ETS term for paragraphs) in an essay and their average length (Attali & Burstein, 2006; Attali & Powers, 2008; Quinlan, Higgins, & Wolff, 2009). Criterion, the classroom adaptation of e-rater, flags any paragraph with fewer than four sentences.* [Perelman, 2020]

Ein solcher Ansatz funktioniert als Annäherung nur so lange gut, wie die Schreibenden die Art und Weise, wie sie einen solchen Text verfassen, nicht an die neuen Bedingungen anpassen. An zwei Enden werden die Defizite dieser Quantifizierung menschlichen Schreibens deutlich: Zum Einen würde ein solches automatische Essaybewertungssystem Schreibende schlecht bewerten, die es schaffen, originelle Gedanken und Argumente in kurze und einfache Sätze zu verpacken – obwohl das eine von uns Menschen hochgeschätzte und seltene Fähigkeit ist. Das oben zitierte Beispiel der Ansprache von Gettysburg zeigt das Problem deutlich. Zum anderen können Prüflinge und Dritte, die die Kriterien des angewendeten Systems einmal durchschaut haben, dieses leicht ausspielen: So erschuf beispielsweise der MIT-Wissenschaftler Les Perelman einen Algorithmus namens *Basic Automatic BS Essay Language Generator* (BABEL)[1], der zwar die vom System erwarteten Überprüfungskriterien berücksichtigt, inhaltlich aber völlig sinnbefreite Texte generiert. Die Abkürzung BS im Akronym BABEL steht daher zweifelsfrei für *bullshit*, für „völligen Unsinn“.

[1] Der Generator ist hier online verfügbar: *https://babel-generator.herokuapp.com/* .

Die von BABEL generierten Texte erzielten anschließend beste Bewertungen in verschiedenen automatischen Essaybewertungssystemen, die heute flächendeckend in den USA verwendet werden. Obwohl diese Tests für die Aufnahme an begehrten Universitäten so ausschlaggebend sind, zeigen diese Ergebnisse klar, wie einfach es ist, die automatischen Essaybewertungssysteme auszutricken. In seinem Forbes-Artikel „No, Software Still Can't Grade Student Essays" bringt Peter Greene genau diese Veränderung der Anreizstruktur für Prüflinge auf den Punkt:

> *The ultimate argument about Perelman's work with BABEL is that his submissions are "bad faith writing." That may be, but the use of robo-scoring is bad faith assessment. What does it even mean to tell a student, "You must make a good faith attempt to communicate ideas and arguments to a piece of software that will not understand any of them."* [Greene, 2020]

In der Folge entstand ein fast schon bizarres Wettrüsten, indem die verantwortliche Organisation hinter dem GRE-Test, die *Educational Testing Services* (ETS), wiederum einen Algorithmus konzipierte, um BABEL-Texte zu erkennen [Perelman, 2020]. Ob diese Mechanismen aber auch die auf die Software zielenden Manipulationen menschlicher Kandidat:innen erkennen können, ist unklar.

Solche Reaktionen und Gegenreaktionen sind typisch für **sozioinformatische Systeme**, also solche Systeme, die aus einem oder mehreren sozialen Systemen bestehen, die über eine Software miteinander interagieren. Software wie die beschriebenen automatischen Essaybewertungssysteme verändert dabei die Anreizstrukturen der beteiligten Personen und Institutionen: Für die Institutionen werden die Gutachten auf lange Sicht günstiger, die Kandidat:innen bemerken aber gleichzeitig, dass ein paar einfache Tricks ihnen zu besseren Noten verhelfen können. Damit wird die Bewertung weniger aussagekräftig und unter Umständen müssen die Universitäten als Nutzer der Bewertung dann wiederum mehr Aufwand in die Bewertung von Bewerbungsunterlagen stecken, um eine ertrickste Bestleistung zu identifizieren. Dieser gesamtgesellschaftliche Effekt kann auf den ersten Blick als überraschend eingestuft werden.

Tatsächlich sind die Reaktionen und Gegenreaktionen des sozialen Systems auf die Einführung eines Softwaresystems **aus der Perspektive der durch diesen Einsatz veränderten Anreizstrukturen** weniger überraschend. Eine Definition des Begriffs der Anreizstruktur findet sich in Abschnitt 4.6, grob verstehen wir darunter die Menge der Regeln und Prozesse in einem sozialen oder sozioinformatischen System, die bestimmen, welches menschliche Verhalten zu welchem Ergebnis beiträgt. In diesem Fall bietet der Einsatz der Software für die Verwender:innen unter anderem den Anreiz, Kosten für menschliche Gutachtachter:innen einzusparen, und setzt gleichzeitig den Kandidat:innen Anreize, möglichst lange Texte zu schreiben – mit möglichst vielen selten genutzten Wörtern.