Tim Breuer, Susanne Keil

Visuelle Darstellungen von MINT-Berufen durch Bildgeneratoren. Wie viel Vielfalt ist möglich?

Prof. Dr. Susanne Keil (Foto: Bettina Steinacker).

1 Einleitung

In den vergangenen Jahren haben sich Text-zu-Bild-Transformer-Modelle wie DALL·E, Stable Diffusion und Midjourney etabliert, die realitätsnahe Bilder generieren. So wurden zwischen 2022 und 2023 über 15 Milliarden KI-Bilder produziert, Midjourney alleine zeigt eine Nutzendenbasis von 16 Millionen (Broz 2023; Valyaeva 2023; Zhou et al. 2024). Diese kritische retrospektive Analyse beschäftigt sich mit DALL·E Mini, einem der ersten öffentlich weit verbreiteten schwächeren Modelle, das für viele Nutzende den initialen Kontaktpunkt mit dieser Technologie darstellte. Im Sommer 2022 erreichte DALL·E Mini in den sozialen Medien seinen Viralitätshöhepunkt (O'Meara/Murphy 2023: 1071f.), als Nutzende surreale Bildkompositionen auf Plattformen wie Twitter und Reddit teilten, was zur massiven Verbreitung der Technologie beitrug (O'Meara/Murphy 2023: 1072).

Frühzeitig wurden jedoch auch algorithmische Verzerrungen (Bias) kritisiert. Journalistische Artikel wie „The AI That Draws What You Type Is Very Racist Shocking No One" (Rose 2022) beleuchteten, wie das Modell Stereotypisierung reproduzierte (vgl. Rose 2022). Wissenschaftliche Untersuchungen bestätigen ebenfalls Verzerrungen in Bezug auf Geschlecht und Ethnizität bei beruflichen Rollen (vgl. Cheong et al. 2023: 4; Naik/Nushi 2023). Insbesondere bei MINT-Berufen, in denen Frauen und People of Color bereits marginalisiert sind (Anger et al. 2021: 30; Dodiya et al. 2022; Keil/Orth 2023: 31), könnte die algorithmische Reproduktion solcher Stereotypen durch generative Modelle zu einer weiteren Diskriminierung beitragen. Die vorliegende Untersuchung analysiert die von DALL·E Mini generierten Visualisierungen von MINT-Berufen unter der Frage, inwiefern diese Modelle Geschlechter- und Ethnie-Disparitäten zementieren. Darüber hinaus wird das Potenzial ethischer Interventionen durch „Prompt Expansion" beleuchtet.

2 Technische und theoretische Grundlagen

Im folgenden Abschnitt wird die Modellarchitektur von DALL·E Mini beschrieben sowie die Untersuchungsdimensionen „Gender(attribution)" und „People of Color" mit Bezug auf algorithmische Verzerrungen zugespitzt.

2.1 DALL·E Mini: Technologische Grundlagen

DALL·E Mini stellt ein auf Transformer-Architekturen basierendes Modell zur Text-zu-Bild-Generierung dar, das darauf abzielt, Darstellungen auf der Grundlage semantischer Texteingaben zu synthetisieren. Es wurde als Open-Source-Pendant zu DALL·E konzipiert, um den Zugang zu hochentwickelten Text-Bild-Modellen zu demokratisieren (Dayma et al. 2021, 2022). Das Modell nutzt neuronale Netzwerke, die auf umfangreichen, weitgehend unbeaufsichtigten Datensätzen trainiert wurden, darunter Conceptual Captions mit drei Millionen Bild-Text-Paaren (z. B. Stockfotos mit korrespondierenden Alternativtexten), Conceptual 12M mit zwölf Millionen Paaren und eine auf zwei Millionen Paare reduzierte Version von YFCC100M (Dayma et al. 2022; Sharma et al. 2018; Changpinyo et al. 2021; Thomee et al. 2016). Der VQGAN-Encoder (Vector Quantized Generative Adversarial Network) quantisiert die Bilddaten in latente

Tim Breuer (Foto: Bettina Steinacker).

(niedrigdimensional vereinfachte) Vektordarstellungen, die durch diskrete Token (kleine Einheiten) codiert werden (Dayma et al. 2022). Parallel verarbeitet der BART-Encoder (Bidirectional and Auto-Regressive Transformer) die Textinformationen, die die Bildbeschreibungen enthalten, und der BART-Decoder generiert sequentiell Bildtoken basierend auf diesen kodierten Repräsentationen (Dayma et al. 2022; Lewis et al. 2019; Raval 2023). Die Optimierung erfolgt während der Trainingsphase durch Minimierung der Diskrepanz zwischen den prognostizierten und den tatsächlichen Bilddaten mithilfe der Softmax-Cross-Entropy-Loss-Funktion (vgl. Dayma et al. 2022; Maheshkar 2023). Insgesamt dient die Trainingspipeline einer kohärenten Verknüpfung zwischen Text- und Bilddimension, während bei der Inferenzpipeline Texteingaben von Nutzenden durch BART (Seq2Seq „trained/fine-tuned") zu Samples von Bildkodierungen übersetzt werden, wobei das neuronale Netzwerk „CLIP" voraussichtlich passende Ergebnisse (Sample) auswählt (Dayma et al. 2022).

2.2 Gender(attribution) bei synthetischen Figuren

Die Auseinandersetzung mit Gender(attribution) in synthetischen Bildkreationen eröffnet einen Diskurs über die Konstruktion und Zuschreibung von Geschlecht in der digitalen Sphäre. Zentral für dieses Verständnis war Judith Butlers Theorie der Performativität, die Geschlecht nicht als inhärente determinierte Eigenschaft, sondern als ein durch wiederholte gesellschaftliche Praktiken (Sprachakt) und Diskurse konstituiertes Konstrukt begreift (vgl. Butler 1999: 43). Diese Macht des Diskurses (Butler 1995: 22) erzeugt Phantasmen (Trugbild eines natürlichen Geschlechts), normative Vorstellungen, die als kulturelles Verständnis dienen (Butler 1995: 23). Butler argumentiert, dass der „Körper als Ort der Möglichkeiten" selbst als konstruiert betrachtet werden sollte, geformt durch Zwänge und Einschränkungen, die einer kulturellen Begreiflichkeit dienen (Butler 1995: 11, 28). Im Kontext von DALL·E Mini soll folglich die Geschlechtszuordnung von generierten Figuren in generativen Bildern rein durch eine Genderattribution geschehen, die auf performativen Praktiken (verbale Zuordnung) und normativen Phantasmen basiert („congealing of the body over time"), um sich von dem Begriff „Identität" (als „rigid inner truth") zu entfernen. Hier können der genderkonforme Ausdruck und das endogeschlechtliche Körperverständnis (vgl. Hübscher 2016; 2022) zu analytischen Zwecken herangezogen werden.
Wie bereits festgestellt, lassen sich konkrete Aussagen über die Geschlechtsidentität anhand der generativen Bilder nicht treffen, lediglich eine Zuordnung. Die Benennung von Ausdruck und Körper bei generativen Bildern könnte mit Bezug auf Butler als eine zyklische Gewalt des Dialogs interpretiert werden, die jene Geschehnisse hervorbringt, die sie zu regeln und einzugrenzen versucht (Butler 1995: 22). Die Kategorisierung von generativen Bildern könnte die normativen Vorstellungen von Geschlecht als symbolische Formen wiederholen oder sogar infrage stellen. Generative Bilder von nichtexistierenden Personen (Figuren) bieten ein besonders gutes Beispiel für die rituelle Geschlechtszuschreibung, da sich die generierten Personen nicht nach ihrer Geschlechtsidentität fragen lassen. In der folgenden Inhaltsanalyse werden die Figuren bei der Kodierung von Gender daher mit dem Suffix „-konstruiert" bezeichnet, um deutlich zu machen, dass es sich um eine Zuweisung durch eine kulturelle Matrix handelt.

2.3 People of Color und Automated Anti-Blackness

„Race" ist, wie Naik und Nushi (2023) betonen, ein soziales Konstrukt, das auf physischen Merkmalen wie Hautfarbe, Haarstruktur und Gesichtszügen basiert, während Ethnizität auf kulturelle Unterschiede verweist. Der Begriff „People of Color" bezieht sich auf Gruppen, die durch Rassismus-Erfahrungen ausgegrenzt werden (Broden/Mecheril 2014; Ha 2019; Safire 1998). Im Bereich des Machine Learnings manifestieren sich diese sozialen Konstruktionen in algorithmischen Verzerrungen, die People of Color benachteiligen. So führten etwa Suchanfragen mit Namen, die häufig bei Schwarzen Personen vorkommen, vermehrt zu negativen Suchergebnissen wie Anzeigen über Verhaftungen (Sweeney 2013). Weitere Beispiele sind

medizinische Datenbanken, die überwiegend europäische Genmaterialien nutzen, was zu Fehldiagnosen bei nicht-weißen Patient*innen führen kann (Popejoy/Fullerton 2016). Besonders problematisch ist die Gesichtserkennungssoftware, die bei Schwarzen Frauen die höchste Fehlerrate aufweist (Buolamwini/Gebru 2018). Fälle wie die ungerechtfertigte Festnahme von Robert Williams zeigen die weitreichenden Folgen solcher Datenlücken in realen Kontexten (Bhuiyan 2023). Mutale Nkonde beschreibt diese systemischen Verzerrungen als „automated anti-blackness" (Nkonde 2019; Rauenzahn et al. 2021). Von Relevanz ist, ob und wie sich diese Muster in generativer KI weitertragen.

3 Forschungsstand

Bevor der Bias in generativen KI-Modellen untersucht wird, ist eine Betrachtung des verwandten Bereichs des Machine Learning geboten. Das Programm „Gendered Innovations" der Stanford University hebt hervor, dass Künstliche Intelligenz (KI) implizit historische Vorurteile in zukünftige Anwendungen übertragen kann (Gendered Innovations: Machine Learning: Analyzing Gender). Diese Problematik wird in zahlreichen Studien bestätigt: So zeigt eine Untersuchung von Datta et al. (2015), dass in Google-Suchanzeigen für Führungspositionen Männer fünfmal häufiger ausgespielt werden als Frauen. Ähnliche algorithmische Verzerrungen zeigten sich auch bei Google Translate, wo männliche Pronomen als Default verwendet wurden, selbst wenn der Ursprungstext explizit auf eine Frau verwies (Minkov et al. 2007). Bolukbasi et al. (2016) demonstrierten, dass Word Embeddings, die semantische Beziehungen zwischen Begriffen lernen, „computer programmer" systematisch mit „man" und „housemaker" mit „woman" assoziierten.

Eine interne Evaluation von DALL·E Mini durch Dayma et al. (2021; 2022) ergab, dass das Modell überwiegend weiß-kodierte Figuren generiert, die gegenüber People of Color häufig Machtpositionen einnehmen. Hochqualifizierte Berufe oder Berufe mit physischer Anstrengung wie „Engineer" oder „Construction Worker" werden fast ausschließlich mit weiß- und männlich-kodierten Figuren verknüpft. Diese Verzerrungen lassen sich auf die zugrunde liegenden Datensätze und die Architektur zurückführen.

Die Untersuchung von Cheong et al. (2023) betont die fundamentale Bedeutung der Trainingsdaten, da generative Text-zu-Bild-Modelle auf den in den Datensätzen manifestierten sozialen Bias zurückgreifen, um visuelle Relationen zu erzeugen (Cheong et al. 2023: 1). Cheong et al. (2023: 3) nutzten Berufsbezeichnungen in DALL·E Mini, um die generierten Figuren hinsichtlich ihres „perceived gender" (Mann, Frau, undefiniert) und ihrer „perceived racial identity" (weiß oder nicht-weiß) zu klassifizieren. Diese Ergebnisse wurden mit der realen Berufspopulation verglichen, basierend auf Daten des U. S. Bureau of Labor Statistics. Dabei stellte sich eine bimodale Verteilung heraus – es wurden fast ausschließlich männlich- oder weiblich-kodierte Figuren generiert –, die deutlich von der realen Verteilung abweicht. Diese Befunde legen nahe, dass generative KI nicht nur bestehende Stereotype reproduziert, sondern auch verstärkt. Berufe wie „Waiter", „Baker", „Accountant", „Biologist", „Poet" und „Judge", die laut den Labor-Statistiken einen hohen Frauenanteil haben, wurden in den DALL·E Mini-Bildern hingegen durch einen niedrigeren Anteil weiblich-kodierter Figuren dargestellt (Cheong et al. 2023: 4). Dass dies in unterschiedlicher Ausprägung auch für andere Text-to-Image-Generatoren gilt, belegen Naik und Nushi (2023: 1). Es zeigt sich der Trend einer doppelten Verzerrung: zuerst durch die Trainingsdaten, dann durch die gelernten Relationen, die die Generierung beeinflussen. Um diesen Verzerrungen entgegenzuwirken, wurden im Bereich des sogenannten „Prompt Engineering" bereits Ansätze entwickelt. Bansal et al. (2022) zeigten, dass ethische Sprachinterventionen, wie die Methode „ENTIGEN" („Ethical NaTural Language Interventions in Text-to-Image GENeration"), die Repräsentation von Geschlechtern und ethnischen Gruppen in generativen Modellen durch Prompt-Anweisungen wie „if all individuals can be a lawyer irrespective of their gender" diversifizieren können, ohne dabei explizit auf Personengruppen hinzuweisen.

4 Methode

Die qualitative Inhaltsanalyse fokussierte sich auf die Untersuchung von Diversitätsdimensionen in MINT-Berufen. Zur Vermeidung von grammatikalisch bedingten Geschlechtsmarkierungen, wie sie im Deutschen vorherrschen, wurden zwanzig Textanweisungen (Prompts) auf Englisch formuliert. Jeder Beruf-Prompt generiert bei DALL·E Mini neun Bilder. Die Prompts sind strukturiert in vier Disziplinen: Mathematik, Informatik, Naturwissenschaften und Technik (MINT). Insgesamt resultierten hieraus 180 Berufsdarstellungen, aufgeteilt in jeweils fünf Berufe pro Disziplin. Die Genderattribution analysiert Geschlechterdarstellungen und endogeschlechtliche Körperkons-

truktionen. Hier stand die visuelle Manifestation kulturell normativer Geschlechtervorstellungen im Vordergrund, die sich etwa durch spezifische Marker wie Kleidung, Frisur, Styling, Körperform und Körpersprache konkretisierten. Die endogeschlechtliche Körperwahrnehmung thematisiert die normativ-binäre Vorstellung von Körpern, die spätestens bei der Geburt fremd zugeordnet wird (Sprechakt). Obwohl die Forschung die Haltbarkeit dieser binären Sicht längst widerlegt hat, bleibt sie in der Gesellschaft tief verankert (vgl. Hübscher 2022). Die zweite Kategorie umfasste die People of Color (PoC), wobei „Race" als konstruiertes soziales Phänomen verstanden wurde, das auf sichtbaren physischen Merkmalen wie Hautfarbe und Gesichtszügen basiert, während „Ethnizität" den kulturellen Hintergrund und historische Identitätsaspekte einer Person beschreibt (vgl. Naik/Nushi 2023). Nach der Kodierung wurden die Berufe mit geschlechts- und ethniedemografischen Daten zu den jeweiligen Belegschaften verglichen, um potenzielle Verzerrungen greifbarer zu machen. Hierauf aufbauend wurde Prompt-Engineering mit Lenkung durchgeführt, um in Abgrenzung zur Studie von Bansal et al. (2022) gezielt die Darstellung von Geschlechter- und Ethnizitäts-Vielfalt zu beeinflussen. Diese ethische Intervention, durchgeführt anhand eines Berufs aus jeder MINT-Kategorie, soll untersuchen, wie DALL·E Mini Anweisungen, die sich direkt auf Geschlecht („female") und Ethnizität („black") verweisen, umsetzt.

5 Ergebnisse

Es folgen zunächst die demografischen Daten zu den Belegschaften der zwanzig MINT-Berufe, um eine Gegenüberstellung mit den kodierten Bildern durch DALL·E Mini zu ermöglichen.

5.1 Gender(attribution)

Im Segment der mathematischen Berufe (M-Berufe), etwa „Statistician" und „Business Mathematician", beträgt der reale Frauenanteil zwischen 36.8 % und 40.8 %, wohingegen die generative KI Frauen gänzlich unsichtbar macht (0 %). Lediglich für „Math Teacher" nähert sich DALL·E Mini der Realität an, mit 44.4 % Frauenanteil statt 53 %.
In den Informatikberufen (I-Berufe) wie „Software Engineer", „Computer Scientist" und „IT Project Manager" liegt der Frauenanteil in der Realität zwischen 20.2 und 33.6 %, doch DALL·E Mini generiert ausschließlich männlich-konstruierte Darstellungen (0 %), was auf eine

Tabelle 1: Gegenüberstellung Frauenanteil der Belegschaften und weiblich-konstruierte Figuren durch DALL·E Mini

M-Berufe	Statistik	DALL E	I-Berufe	Statistik	DALL E
Statistician (Worker)	36.8 %	0	Software Engineer	20.2 %	0
Accountant	0,57	33.3 %	Computer Scientist	21.2 %	0
Supply Chain Manager	21.4 %	33.3 %	IT Project Manager	33.6 %	0
Business Math	40.8 %	0	Web Developer	19.4 %	0
Math Teacher	0,53	44.4 %	System Admin	16.7 %	0
N-Berufe	**Statistik**	**DALL E**	**T-Berufe**	**Statistik**	**DALL E**
Biologist	0,55	55.5 %	Engineer	13.7 %	0
Chemist	0,36	55.5 %	Electrician	2.9 %	0
Physicist	16.1 %	0	Mechanical Engineer	10.1 %	0
Nurse	87.4 %	1	Architect	23.3 %	0
Pharmacist	57.8 %	66.6 %	Industrial Technician	13.3 %	0

Quelle: eigene Darstellung. Vergleichsdaten von U. S. Bureau of Labor Statistics (2023) und Zippia Demographic Research (2024).

Tabelle 2: Gegenüberstellung PoC-Anteil der Belegschaften und PoC-konstruierte Figuren durch DALL·E Mini

M-Berufe	Statistik	DALL E	I-Berufe	Statistik	DALL E
Statistician (Worker)	40.6 %	0	Software Engineer	45.4 %	11.1 %
Accountant	26.6 %	0	Computer Scientist	35.8 %	0
Supply Chain Manager	34.1 %	0	IT Project Manager	32.3 %	0
Business Math	37.6 %	0	Web Developer	15.8 %	0
Math Teacher	0,28	0	System Admin	26.8 %	0
N-Berufe	**Statistik**	**DALL E**	**T-Berufe**	**Statistik**	**DALL E**
Biologist	15.4 %	22.2 %	Engineer	32.1 %	0
Chemist	35.3 %	22.2 %	Electrician	12.7 %	0
Physicist	26.6 %	0	Mechanical Engineer	21.1 %	0
Nurse	27.4 %	22.2 %	Architect	34.4 %	0
Pharmacist	31.5 %	0	Industrial Technician	33.2 %	0

Quelle: eigene Darstellung. Vergleichsdaten von U. S. Bureau of Labor Statistics (2023) und Zippia Demographic Research (2024).

drastische Verzerrung hinweist und die männliche Konnotation dieser Berufe verfestigt.

Im Bereich der Naturwissenschaften (N-Berufe) gelingt DALL-E Mini partiell eine Annäherung. So reflektiert das Modell Berufe wie „Biologist" oder „Pharmacist" mit Frauenanteilen von 55.5 % und 57.8 % relativ nahe an den realen 55 % bzw. 66.6 %. Dies überrascht nicht, da die realen Frauenanteile höher sind und Mädchen ein höheres Vertrauen in ihr Vorwissen in Biologie äußern (Dodiya et al. 2022: 2). Zudem belegten Studien bereits ein stärkeres Interesse von Mädchen an MINT-Studiengängen mit Umwelt- und Naturschutzaspekten (Mohaupt 2017: 11). „Physicist" sticht mit einer ausschließlich männlich-konstruierten Visualisierung hervor. DALL·E Mini stellt allerdings nur weiblich-konstruierte Figuren für „Nurse" dar, ein Care-Beruf. So manifestieren sich Stereotypisierungen nicht ausschließlich durch Unter-, sondern auch durch Überrepräsentation.

Technische Berufe (T-Berufe) wie „Engineer", „Electrician" und „Industrial Technician" zeigen eine Verstärkung bestehender Marginalisierung: Trotz eines ohnehin niedrigen Frauenanteils (2.9 % bis 13.3 %) stellt DALL·E Mini ausschließlich männlich-konstruierte Figuren für den Techniksektor dar. Insgesamt offenbart DALL·E Mini eine systematische Unsichtbarmachung von Frauen in MINT-Berufen (insgesamt 145 von 180 Bilder zeigen keine weiblich-konstruierten Figuren), vor allem in den bereits männlich konnotierten Bereichen Informatik und Technik. Die partiell realitätsnahe Darstellung in naturwissenschaftlichen Berufen bleibt eine Ausnahme.

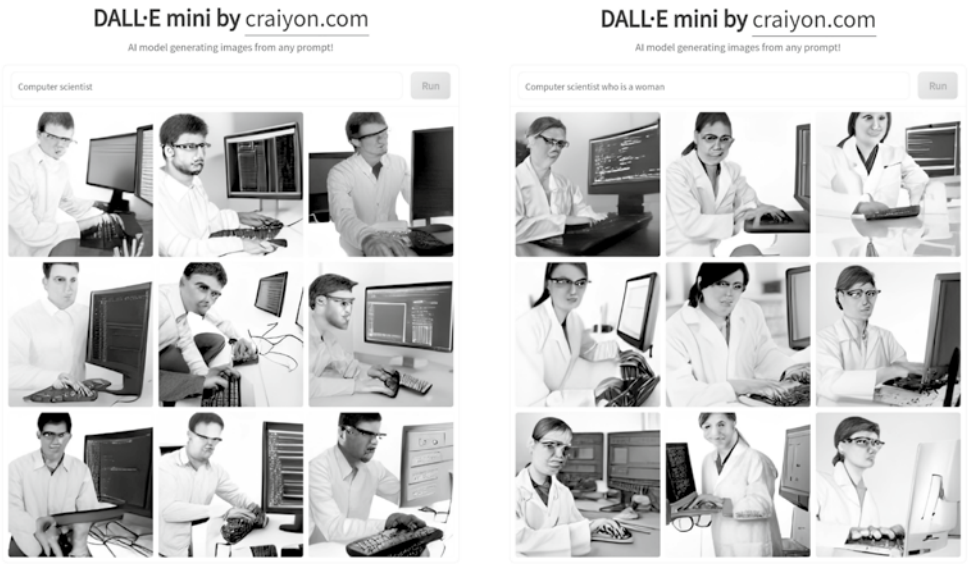
5.2 PoC-Anteil

In der Kategorie Mathematik, zu der Berufe wie „Statistician", „Supply Chain Manager" und „Math Teacher" zählen, variiert der statistische Anteil von People of Color (PoC) zwischen 26.6 % und 40.6 %. DALL·E Mini jedoch eliminiert PoC gänzlich aus diesen Berufsbildern (0 %). Auch im Bereich Informatik, der Berufe wie „Software Engineer", „IT Project Manager" und „Web Developer" umfasst, wird die statistische Präsenz von PoC kaum berücksichtigt. Trotz realer Anteile zwischen 15.8 % und 45.4 % erscheinen PoC in diesen Darstellungen nahezu gar nicht, mit maximal 11.1 % bei „Software Engineer" und 0 % in den anderen Informatikberufen. Eine der Realität angemessene Konstruktion zeigt sich in den Naturwissenschaften, etwa bei „Biologist", „Chemist" und „Nurse". „Physicist" und „Pharmacist" bilden jedoch wieder keine PoC ab.

Abbildung 1: Gegenüberstellung Prompt „Engineer" und ethische Intervention „Black female engineer"

Quelle: eigene Darstellung mit DALL·E Mini.

Abbildung 2: Gegenüberstellung Prompt „Computer Scientist" und ethische Intervention „Computer scientist who is a woman"



Quelle: eigene Darstellung mit DALL·E Mini.

Der technische Bereich weist eine markante Diskrepanz auf: Obwohl der PoC-Anteil statistisch bis zu 34.4 % erreicht, bleibt die DALL·E Mini-Darstellung von PoC in sämtlichen Technikberufen vollständig aus (0 %). Insgesamt zeigen nur ein informatischer und drei naturwissenschaftliche Berufe Figuren, die als PoC kodiert werden konnten (4 von 20 Berufen und insgesamt 7 von 180 Bildern). Hinsichtlich der Genderattribution gibt es zudem Berufe (z. B. alle technischen Berufe) mit intersektionaler Unsichtbarkeit.

5.3 Ethische Intervention

Die Erweiterung von „Engineer" durch „Black female" führt zu einer ausschließlich weiblich- und PoC-konstruierten Repräsentation, die sich von der rein männlich- und weiß-konstruierten Darstellung ohne Intervention abhebt (siehe Abb. 1). Die Bilder mit ethischer Intervention beim Prompten wirken oft posierend (lächelnd zu Betrachtenden gewendet) und erinnern in ihrer Bildkomposition an Stockfotos aus Bilddatenbanken. Es handelt sich nicht um Fotos

von arbeitenden People of Color, die ihre tatsächliche Kompetenz als aktive Akteur*innen in den Vordergrund stellen. Das Prompt Engineering für „Person of Color who is an accountant" und „Female physicist" zeigen vergleichbare Resultate.
Eine Ausnahme bietet das Prompt Engineering „Computer Scientist who is a woman". Hier verbindet DALL·E Mini wortwörtlich Wissenschaftlerinnen (Scientist who is a woman: weiße Kittel, Schutzbrille) mit Computern, aber generiert keine Informatikerinnen (siehe Abb. 2). Bilder von Informatikerinnen, aus denen DALL·E Mini ziehen kann, sowie erlernte Wortbeziehungen können hierfür verantwortlich sein.

6 Diskussion

Die Überrepräsentation von weiß- und männlichkonstruierten Figuren im MINT-Sektor verdeutlicht eine systemische Verzerrung innerhalb der KI-gestützten Bildgenerierung durch DALL·E Mini, die insbesondere durch die Darstellung in IT- und Technik-Berufen manifest wird: 145 von 180 MINT-Berufsbildern konnten ausschließlich als männlich-konstruiert kodiert werden, während 173 von 180 keine People of Color (PoC) repräsentieren. Auffällig ist die völlige Abwesenheit weiblich-konstruierter Figuren in mathematischen und Informatik-Berufsprofilen. Der Abgleich mit tatsächlichen Beschäftigungsdaten untermauert die verstärkende Rolle generativer KI bei der Reproduktion existierender Ungleichheiten. Der perpetuierte Bias in technischen Anwendungen ist das Resultat menschlicher Entscheidungsmuster, insbesondere aufseiten der Entwickelnden. Die enge Verknüpfung zwischen KI und Gesellschaft (Vlasceanu/Amodio 2022) zeigt sich hier deutlich in algorithmischen Systemen, die Machtstrukturen nicht nur reflektieren, sondern auch reproduzieren. Die Ergebnisse dieser Untersuchung reihen sich in bestehende Studien (Cheong et al. 2023; Naik/Nushi 2023) ein und konkretisieren einen Bias generativer Modelle auf den MINT-Sektor als aussagekräftiges Beispiel. Dies verdeutlicht die Notwendigkeit eines reflektierten Umgangs mit KI, Automatisierung und maschinellem Lernen. Die Studienergebnisse demonstrieren, dass mittels Prompt-Expansion eine partielle Ausbalancierung algorithmischer Exklusion möglich ist, die strukturellen Ursachen jedoch nicht beseitigt werden. Die Hauptverantwortung für ethisches Monitoring sollte bei den Entwickelnden liegen: Eine kontinuierliche Einbindung („Mobilisierungsdiskurs") von heterogenen Ethikteams aus den Geisteswissenschaften in der gesamten Entwicklungspipeline wäre essenziell, um einen Beitrag zu fairen und inklusiven Technologien zu leisten. Wie Cheong et al. (vgl. 2023: 1) bereits festgestellt haben, lohnt es sich, zu hinterfragen, ob technische Produkte mit Bias überhaupt für die Gesellschaft verfügbar gemacht werden dürfen. Journalistische Artikel, die Bildtransformer wie DALL·E Mini bereits als potenziell sexistisch und rassistisch identifizierten (vgl. Rose 2022; vgl. Al-Sibai 2022), formen den öffentlichen Diskurs und informieren Frauen und People of Color darüber, dass die Technik sie „nicht willkommen" heißt. Dies könnte den Voreingenommenheitskreislauf weiterführen, sodass sich der Techniksektor durch seine diskriminierenden Produkte und Arbeitskultur selbst isoliert und hegemonialisiert (Beispiel Fachkräftemangel).
Die Ergebnisse dieser Studie verdeutlichen die Notwendigkeit, KI als ein von menschlicher Prägung beeinflusstes Werkzeug zu betrachten. Die vorliegende Arbeit zeigt eine Duldung von diskriminierender Technik als „Kompromiss" im Wettrennen um OpenAI-Profit. So zeigen auch aktuelle Bemühungen, generative KI zu „detoxifizieren", durch MetaAI und Google Gemini diese hastige Natur: Wo vorab nur weiß-konstruierte Figuren generiert wurden, zeigen diese Modelle nun Schwarze Personen als Kolonialisierer, amerikanische Gründerväter und Päpste durch automatisierte ethische Intervention – auch wenn diese keinen Sinn ergibt (Gilbert 2024; Shamim 2024; Morrone 2024). Rechte Akteur*innen wussten diesen „historischen Revisionismus" direkt für ihren Kulturkampf zu nutzen (siehe „Woke AI", Gilbert 2024). Unverantwortliche Unterrepräsentation und unausgefeilte Überkompensation schaden letztlich beide marginalisierten Gruppen und zeigen, dass generative KI nach wie vor Frauen und People of Color nicht zuverlässig, respektvoll und realistisch darstellen kann.

Literaturverzeichnis

- Al-Sibai, Noor (2022). That AI Image Generator Is Spitting Out Some Awfully Racist Stuff. *Futurism.* Zugriff am 06. November 2024 unter https://futurism.com/dall-e-mini-racist.
- Anger, Christina et al. (2021). MINT-Herbstreport 2021: Mehr Frauen für MINT gewinnen – Herausforderungen von Dekarbonisierung, Digitalisierung und Demografie meistern. *Institut der deutschen Wirtschaft Köln e. V. (Hrsg.),* S. 33–36.
- Bansal, Hritik et al. (2022). How well can Text-to-Image Generative Models understand Ethical Natural Language Interventions?

Conference on Empirical Methods in Natural Language Processing. Zugriff am 06. November 2024 unter https://arxiv.org/abs/2210.15230.
- Bhuiyan, Johana (2023). First man wrongfully arrested because of facial recognition testifies as California weighs new bills. *The Guardian.* Zugriff am 06. November 2024 unter www.theguardian.com/us-news/2023/apr/27/california-police-facial-recognition-software.
- Birhane, Abeba et al. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *arXiv preprint,* arXiv:2110.01963.
- Bolukbasi, Tolga et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems,* S. 4349–4357.
- Broden, Anne & Mecheril, Paul (2014). *Rassismus bildet: bildungswissenschaftliche Beiträge zu Normalisierung und Subjektivierung in der Migrationsgesellschaft.* transcript Verlag, S. 144ff.
- Broz, Matic (2023). Midjourney statistics (November 2024). *Photutorial.* Zugriff am 06. November 2024 unter https://photutorial.com/midjourney-statistics/.
- Buolamwini, Joy & Gebru, Timnit (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Conference on Fairness, Accountability and Transparency,* S. 77–91.
- Butler, Judith (1995). *Körper von Gewicht: Die diskursiven Grenzen des Geschlechts.* Aus dem Amerikanischen von Karin Wördemann. Berlin-Verlag, Berlin.
- Butler, Judith (1999). *Gender Trouble: Feminism and the Subversion of Identity.* Routledge, New York.
- Changpinyo, Soravit et al. (2021). Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training to Recognize Long-Tail Visual Concepts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.48550/arXiv.2102.08981.
- Cheong, Marc et al. (2023). Investigating Gender and Racial Biases in DALL-E Mini Images. ACM J, *Responsib. Comput.* 1(2), Artikel 13. https://doi.org/10.1145/3649883.
- Datta, Amid et al. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies,* 1, S. 92–112. https://doi.org/10.1515/popets-2015-0007.
- Dayma, Boris et al. (2021). DALL·E Mini. *Hugging Face.* Zugriff am 06. November 2024 unter https://huggingface.co/dalle-mini/dalle-mini#limitations-and-bias-recommendations.
- Dayma, Boris et al. (2022). DALL-E Mini Explained. *Weights & Biases.* Zugriff am 06. November 2024 unter https://wandb.ai/dalle-mini/dalle-mini/reports/DALL-E-Mini-Explained-with-Demo--Vmlldzo4NjIxODA.
- Dodiya, Janki et al. (2022). Mint-Bildung: Was junge Frauen darüber denken. IU Internationale Hochschule. Zugriff am 06. November 2024 unter https://static.iu.de/studies/Junge_Frauen_in_MINT_Kurzstudie.pdf.
- Gendered Innovations (o. A.). Machine Learning: Analyzing Gender. *Gendered Innovations, Stanford University.* Zugriff am 06. November 2024 unter https://genderedinnovations.stanford.edu/case-studies/machinelearning.html#tabs-2.
- Gilbert, David (2024). Google's 'Woke' Image Generator Shows the Limitations of AI. *Wired.* Zugriff am 06. November 2024 unter www.wired.com/story/google-gemini-woke-ai-image-generation/.
- Ha, Kien Nghi (2019). 'People of Color' als Diversity-Ansatz in der antirassistischen Selbstbenennungs- und Identitätspolitik. *migration-boell.de.*
- Hübscher, Evianne (2016). Grundlagen zum Thema Geschlecht [überarbeitet]. *nonbinary.ch.* Zugriff am 06. November 2024 unter www.nonbinary.ch/grundlagen/.
- Hübscher, Evianne (2022). Normative und expansive Ausprägung von Geschlecht. *Geschlechter-Radar.org.* Zugriff am 06. November 2024 unter https://www.geschlechter-radar.org/normativ-expansiv/.
- Keil, Susanne & Orth, Juliane (2023). Technikvideos für Mädchen? Eine Studie zur Attraktivität von Technik auf YouTube. *Internationales Zentralinstitut für das Jugend- und Bildungsfernsehen (Hrsg.), Televizion: Mädchen und MINT,* Ausgabe 36/2023/1, München: IZI, S. 31–34.
- Lewis, Mike et al. (2019). BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics 2020.* https://doi.org/10.18653/v1/2020.acl-main.703.
- Maheshkar, Saurav (2023). The Softmax Activation Function Explained. *Weights & Biases.* Zugriff am 06. November 2024 unter https://wandb.ai/sauravm/Activation-Functions/reports/Activation-Functions-Softmax--VmlldzoxNDU1Njgy#%F0%9F%93%A2-softmax-+-cross-entropy-loss-(caution:-math-alert).

- Minkov, Einat et al. (2007). Generating Complex Morphology for Machine Translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, S. 23–30.
- Mohaupt, Franziska et al. (2017). MINT the gap – Umweltschutz als Motivation für technische Berufsbiographien? *UBA-Texte 111/2017.* Zugriff am 06. November 2024 unter www.umweltbundesamt.de/sites/default/files/medien/1410/publikationen/2017-12-11_texte_111-2017_mint-the-gap_0.pdf.
- Morrone, Megan (2024). Meta AI creates ahistorical images, like Google Gemini. *Axios.* Zugriff am 06. November 2024 unter www.axios.com/2024/03/01/meta-ai-google-gemini-black-founding-fathers.
- Naik, Ranjita & Nushi, Besmira (2023). Social Biases through the Text-to-Image Generation Lens. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, ACM 2023.* https://doi.org/10.1145/3600211.3604711.
- Nkonde, Mutale (2019). Automated Anti-Blackness: Facial Recognition in Brooklyn, New York. *Harvard Kennedy School Journal of African American Policy,* Vol. 20, S. 31.
- O'Meara, Jennifer & Murphy, Caít (2023). Aberrant AI creations: co-creating surrealist body horror using the DALL-E mini text-to-image generator. *Convergence*, Bd. 29, Nr. 4, S. 1070–1096. Zugriff am 06. November 2024 unter https://doi.org/10.1177/13548565231185865.
- Popejoy, Alice B. & Fullerton, Stephanie M. (2016). Genomics is failing on diversity. *Nature,* 538(7624), S. 161–164. https://doi.org/10.1038/538161a.
- Rauenzahn, Brianna et al. (2021). Facing Bias in Facial Recognition Technology. *The Regulatory Review.* Zugriff am 06. November 2024 unter www.theregreview.org/2021/03/20/saturday-seminar-facing-bias-in-facial-recognition-technology/.
- Raval, Param (2023). Transformers BART Model Explained for Text Summarization. *ProjectPro.* Zugriff am 06. November 2024 unter www.projectpro.io/article/transformers-bart-model-explained/553.
- Rose, Janus (2022). The AI That Draws What You Type Is Very Racist, Shocking No One. *Vice (Motherboard).* Zugriff am 06. November 2024 unter www.vice.com/en/article/wxdawn/the-ai-that-draws-what-you-type-is-very-racist-shocking-no-one.
- Safire, William (1988). On Language; People of Color. *The New York Times,* 20. November 1988.
- Shamin, Sarah (2024). Why Google's AI tool was slammed for showing images of people of colour. *Al Jazeera.* Zugriff am 06. November 2024 unter www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people.
- Sharma, Piyush et al. (2018). Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Melbourne, Australia. Association for Computational Linguistics, S. 556–2565. https://doi.org/10.18653/v1/p18-1238.
- Sweeney, Latanya (2013). Discrimination in online ad delivery. *Queue,* 11(3), S.10. https://doi.org/10.1145/2460276.2460278.
- Thomee, Bart et al. (2016). YFCC100M: The New Data in Multimedia Research. *Communications of the ACM 59(2),* S. 64–73. https://doi.org/10.1145/2812802.
- U. S. Bureau of Labor Statistics (2023). Labor Force Statistics from the Current Population Survey: Employed persons by detailed occupation, sex, race, and Hispanic or Latino ethnicity. *U.S. Bureau of Labor Statistics.* Zugriff am 06. November 2024 unter www.bls.gov/cps/cpsaat11.htm.
- Valyaeva, Alina (2023). People Are Creating an Average of 34 Million Images Per Day. Statistics for 2024. *Everypixel Journal.* Zugriff am 06. November 2024 unter https://journal.everypixel.com/ai-image-statistics.
- Vlasceanu, Madalina & Amodio, David M. (2022). Propagation of societal gender inequality by internet search algorithms. *Proc. Natl. Acad. Sci. U. S. A. 119,* e2204529119. Zugriff am 06. November 2024 unter www.pnas.org/doi/10.1073/pnas.2204529119.
- Zhou, Eric & Lee, Dokyun (2024). Generative artificial intelligence, human creativity, and art. *PNAS Nexus, 3(3),* S. 52. https://doi.org/10.1093/pnasnexus/pgae052.
- Zippia (2024). DEMOGRAPHICS AND STATISTICS IN THE US. *Zippia.* Zugriff am 06. November 2024 unter www.zippia.com/careers/demographics/.

Kontakt und Information

Prof. Dr. Susanne Keil
Professorin für Journalistik.
Schwerpunktprofessur Soziale Nachhaltigkeit und Gender
Prodekanin Fachbereich Ingenieurwissenschaften und Kommunikation
Hochschule Bonn-Rhein-Sieg
Campus Sankt Augustin
Grantham-Allee 20
53757 Sankt Augustin
Tel.: (02241) 865 339
susanne.keil@h-brs.de
www.h-brs.de

https://doi.org/10.17185/duepublico/82759

DOI: 10.17185/duepublico/82759
URN: urn:nbn:de:hbz:465-20241217-083543-4