Bye-bye Bias: Was beim Training generativer KI-Modelle für subjektive Marketingmetriken zu beachten ist

AUTORIN UND AUTOREN

Christina Schamp
Professorin für Marketing, Institut für Digital Marketing & Behavioral Insights, Wirtschaftsuniversität Wien

Jochen Hartmann
Professor für Digital Marketing, TUM School of Management, GenAI Lab, Technische Universität München

Dennis Herhausen
Professor of Marketing, School of Business and Economics, Vrije Universiteit Amsterdam

KEYWORDS

Large Language Models (LLMs), generative KI, Datenqualität, Bias, Marketingmetrik

Von standardisierten zu feinjustierten Large Language Models Die Verbreitung von Large Language Models (LLMs) und multimodalen generativen KI-Systemen wie ChatGPT, Dall-E und Midjourney hat eine neue Ära der Innovation und Exploration eingeläutet. Die Anwendungen versprechen eine breite Palette an Einsatzmöglichkeiten im Marketing – von der Erstellung von Inhalten und personalisierten Empfehlungen über die Generierung von Erkenntnissen bis hin zur internen Prozessoptimierung. LLMs von der Stange weisen für viele Use-Cases oft bemerkenswerte „Zero-Shot"-Fähigkeiten auf und schaffen es, Inhalte zu generieren oder Vorhersagen zu treffen, ohne dass sie explizit auf bestimmte Aufgaben oder den spezifischen Markenkontext trainiert wurden. Diese Modelle werden jedoch mit riesigen Datensätzen aus dem Internet trainiert, wie z. B. Common Crawl oder LAION, die kaum Informationen zur Wahrnehmung von marketingrelevanten Informationen wie dem Markenimage oder dem Customer-Engagement enthalten und den spezifischen Markenkontext nicht berücksichtigen. Es gibt bereits Services, die eine direkte Erstellung von Anzeigen zur Suchoptimierung anbieten. Solche Anzeigen werden für hohe durchschnittliche Klickraten konzipiert und optimiert, bieten aber keine Markendifferenzierung und können im schlimmsten Fall das Markenimage verwässern. Das wahre Potenzial generativer KI (GenAI) wird daher erst durch Feintuning der Modelle für einen spezifischen Markenkontext erschlossen, der sich oft durch eher subjektive Marketingmetriken beschreiben lässt. Im Gegensatz zu objektiven Metriken, die vom Betrachter als „richtig oder falsch" eingestuft werden können (z. B., ob eine Person in einer Anzeige zu sehen ist), erfassen subjektive Metriken Wahrnehmungen, Meinungen, Gefühle oder Überzeugungen zu einer Anzeige, die je nach Betrachter unterschiedlich sein können. Box 1 und Abbildung 1 beschreiben den technischen Prozess des Feintunings eines GenAI-Modells. Die Leistung solcher Modelle hängt wesentlich von der Qualität der zugrunde liegenden Trainingsdaten ab. Daher konzentrieren wir uns darauf, wie man qualitativ hochwertige Trainingsdaten für das Feintuning von GenAI-Modellen erhält. Diese sind entscheidend im Wettbewerb um die besten KI-Anwendungen im Marketing.

»

Mögliche Biases in Trainingsdaten sollten bereits vor dem Training eines KI-Modells erhoben und beseitigt werden.

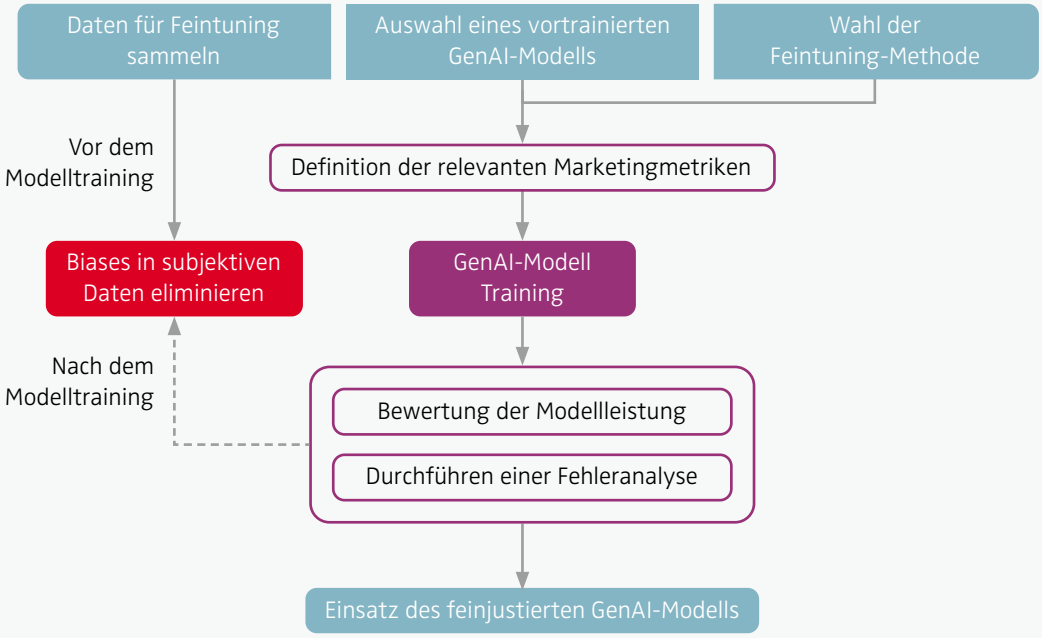
«

BOX 1

Wie Unternehmen GenAI-Modelle für spezifische Marketinganwendungen feinjustieren können

Der Gesamtprozess der Feinabstimmung eines GenAI-Modells ist vergleichbar mit jeder anderen Supervised-Machine-Learning-Pipeline und wird in Abbildung 1 dargestellt. Das verfeinerte Modell, das aus einer solchen mehrstufigen Pipeline hervorgeht, ist eine Funktion von drei Hauptinputs: Trainingsdaten, einem vortrainierten GenAI-Modell wie Llama 2 und einer Feintuning-Methode. Es versteht sich von selbst, dass die Qualität der verwendeten Daten entscheidend ist. Sie kann sowohl vor als auch nach dem Training des Modells bewertet werden. Ex-post-Bewertungen eines Modells und seiner Trainingsdaten werden üblicherweise von Computer- oder Sozialwissenschaftlern durchgeführt und haben das Ziel, Verbesserungsansätze für die Leistung des Modells zu erkennen. Unternehmen sollten jedoch auch Erkenntnisse der Verhaltenswissenschaften nutzen und mögliche Biases bereits vor dem Modelltraining ausschließen. So können durch die repräsentative Auswahl menschlicher Kodierer, durch angemessene Kodierskalen und die Gestaltung einer reibungslosen „Datenannotationsreise" Biases in Trainingsdaten vermieden werden. Diese Faktoren tragen wesentlich zu einer besseren Leistung des verfeinerten GenAI-Modells in realen Anwendungen bei.

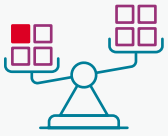
ABBILDUNG 1 > Feintuning-Prozess für GenAI-Modelle



Ein feinjustiertes, vortrainiertes Open-Source-Modell wie Llama 2 ist oft besser als ein Allzweckmodell mit geschlossenem Quellcode wie GPT-3.5. Die konkreten Verbesserungen hängen jedoch vom Anwendungskontext ab. Daher ist es wichtig, den Leistungsunterschied zwischen einem adaptierten GenAI-Modell und einem Basismodell zu messen, bevor es in der Praxis eingesetzt wird. Auch zusätzliche Überlegungen, wie z. B. Datenschutzbedenken, können Unternehmen davon abhalten, eigene Trainingsdaten mit kommerziellen Anbietern zu teilen, was einen Offline-Feintuning-Prozess auf lokalen Servern noch attraktiver macht.

ABBILDUNG 2 › Checkliste zur Sicherung der Datenqualität und zur Bias-Vermeidung in Trainingsdaten

So vermeiden Sie



Sampling-Bias

- Arbeiten Sie mit Zufalls- oder repräsentativen Stichproben, um die Repräsentativität von Bewertungen zu gewährleisten.
- Erheben Sie zusätzliche Daten zu relevanten (demografischen oder segment-spezifischen) Merkmalen und prüfen Sie, ob diese im Training und im realen Leben gleich verteilt sind.
- Überprüfen Sie Non-Response-Bias und ob bestimmte Untergruppen mit geringerer Teilnahmewahrscheinlichkeit ausreichend vertreten sind.
- Gleichen Sie den Datensatz aus, indem Sie relevante unterrepräsentierte Teilgruppen höher gewichten.

Measurement-Bias

- Formulieren Sie kurze, klare und präzise Fragen.
- Verwenden Sie klare Anweisungen und einleitende Kommentare, um Variablen von Interesse bzw. psychologische Konzepte klar zu definieren.
- Nutzen Sie etablierte Skalen, die schwierige Konzepte in mehrere Items/Unterfragen aufschlüsseln, und prüfen Sie die interne Reliabilität der Fragen.
- Wägen Sie ab, ob binäre oder ordinale Skalen genauere Ergebnisse liefern könnten.

Soziale Erwünschtheit – Social-Desirability-Bias

- Betonen Sie in der Einleitung, dass die Daten vertraulich behandelt werden, um ehrliche Antworten zu fördern.
- Formulieren Sie neutrale Fragen, die keine sozialen Normen implizieren.
- Entschärfen Sie sensible Fragen, z. B. durch abgestufte Antwortmöglichkeiten.
- Prüfen Sie die Plausibilität der Antworten anhand früherer Ergebnisse oder vorhandener Daten.

Response-Bias

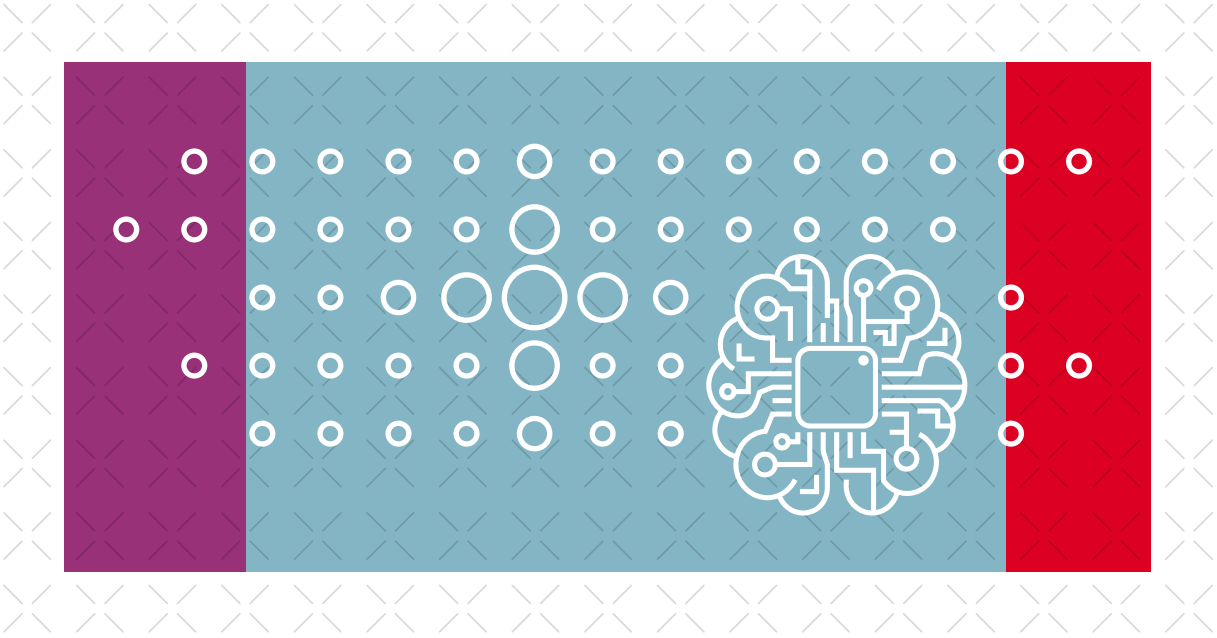
- Bieten Sie ein ansprechendes Befragungsambiente, um Ermüdungstendenzen der Befragten zu minimieren.
- Limitieren Sie die Befragungsdauer und legen Sie die Anzahl der Fragen zu wiederkehrenden Merkmalen entsprechend fest.
- Die Reihenfolge der zu beurteilenden Merkmale sollte nach dem Zufallsprinzip erfolgen, um Reihungseffekte zu vermeiden.
- Filtern Sie Antworten von schlechter Qualität und löschen Sie diese, indem Sie z. B. auf schnelles Klicken, monotone Antwortmuster oder Inkonsistenzen testen.

» *Während bereits objektive Trainingsdaten fehleranfällig sind, ist die Gewährleistung hoher Qualität bei subjektiven Trainingsdaten für Marketingmetriken eine noch größere Herausforderung.* «

Objektive versus subjektive Trainingsaufgaben Bei objektiven Trainingsdaten besteht die zugrunde liegende Aufgabe darin, Elemente zu erkennen oder Objekte als richtig oder falsch zu klassifizieren. In der Regel ist dies objektiv nachvollziehbar: Ist eine Katze abgebildet? Ist ein Logo auf dem Social-Media-Post zu sehen? Ist eine E-Mail Spam? Im Gegensatz dazu erfassen subjektive Trainingsdaten Wahrnehmungen, Meinungen, Gefühle und Überzeugungen, die in verschiedenen Kunden- oder Bevölkerungssegmenten unterschiedlich ausfallen können, z. B. Emotionen, die eine generierte Anzeige hervorruft, ob eine Anzeige lustig ist oder welche Argumente in einer Kundeninteraktion am überzeugendsten sind.

Während bereits objektive Trainingsdaten fehleranfällig sind, ist die Gewährleistung hoher Qualität bei subjektiven Trainingsdaten für Marketingmetriken eine noch größere Herausforderung. Bei objektiven Trainingsdaten reichen meist wenige menschliche Kodierer, um eine geringe Variabilität der Antworten und eine hohe Intercoder-Reliability für die Merkmale zu erreichen. Subjektive Trainingsaufgaben führen hingegen per Definition zu einer höheren Varianz in der Klassifizierung. Diese erfolgt häufig mittels ordinaler oder metrischer Skalen und nicht binär, und im Idealfall wird eine größere und repräsentative Stichprobe von Teilnehmern herangezogen. Konsumentenwahrnehmungen können auch vom geografischen oder kulturellen Kontext abhängen und sich im Laufe der Zeit ändern. Deshalb ist es notwendig, Trainingsverfahren zu wiederholen und Algorithmen regelmäßig zu aktualisieren.

Subjektive Trainingsdaten ermöglichen einen „maßgeschneidert" trainierten Algorithmus, der genau auf die jeweilige Marke, das Produkt oder den Kundenkontext abgestimmt



ist. Um die Qualität der Daten sicherzustellen, müssen Marketer auf Kompetenzen aus unterschiedlichen Bereichen zurückgreifen: verhaltensrelevante Insights aus Befragungen und maschinelles Lernen (ML) auf der Grundlage objektiver Daten. Durch die Einbindung klassischer Marketingerkenntnisse in KI-Modelle lernt GenAI wertvolle Lektionen aus dem traditionellen Marketing: wie man sich von der Konkurrenz abheben, Herz und Hirn der Konsumenten ansprechen und die Rentabilität verbessern kann.

Biases in Trainingsdaten und wie man sie vermeidet Erstaunlich wenig beachtet wurde bislang, wie man zuverlässige, qualitativ hochwertige Trainingsdaten, insbesondere für subjektive Marketingmetriken, erhält. Derzeit wird die Qualität der zugrunde liegenden Daten in erster Linie ex post – nach dem Modelltraining – bewertet, indem systematische Fehler des ML-Modells identifiziert werden. Während sich die Quantifizierung von Fehlern als relativ einfach erweist, ist es schwieriger, auf der Basis von ML-Klassifikationen Fehlerursachen zu verstehen. Erschwerend kommt hinzu, dass Biases in Trainingsdaten systematisch auftreten und deshalb auch im Nachhinein nicht erkennbar sein können.
Wir empfehlen daher, die Qualität der zugrunde liegenden KI-Trainingsdaten nicht erst nach der Anwendung eines ML-Modells zu bewerten, sondern die häufigsten Fallstricke und Biases bei klassischen Marktforschungstechniken wie Umfragen oder Experimenten bereits in einer Vor-Trainings-Phase zu berücksichtigen (siehe Abbildung 1). Die Messung und Verbesserung der Datenqualität sollte also ex ante erfolgen, bevor Daten in ein Trainingsmodell eingespeist werden. Abbildung 2 fasst die wichtigsten Biases in Trainingsdaten sowie entsprechende Empfehlungen zusammen, die man bei subjektiven Marketingmetrik-Aufgabenstellungen beachten sollte, um das Training effizienter zu machen.

- **Sampling-Bias** Dieser Bias tritt auf, wenn Personen, die für das Training des Algorithmus oder die Kennzeichnung der Daten eingesetzt werden, aus einem anderen Umfeld stammen als diejenigen, bei denen der Algorithmus angewendet wird. Dies führt zu systematischen Fehlern in den Modellprognosen. Bei subjektiven Trainingsaufgaben tritt Sampling-Bias auch dann auf, wenn die eingegebenen Daten nicht repräsentativ für die relevante Grundgesamtheit sind und das Stichproben-Sample von der tatsächlichen Verteilung in der relevanten Zielgruppe abweicht. Beispielsweise sollte ein Algorithmus, der messen soll, inwieweit ein Bild die Persönlichkeit einer Marke widerspiegelt, die Wahrnehmungen aller relevanter Kundensegmente beinhalten. Einzelne Kundensegmente können unterschiedliche Wahrnehmungen haben, Märkte können sich kulturell unterscheiden und Konsumentenwahrnehmungen können sich im Laufe der Zeit ändern. Wenn die Trainingsdaten nicht korrekt erfasst werden, kann der Algorithmus keine guten und allgemein gültigen Prognosen liefern.

- **Measurement-Bias** Der Measurement-Bias bezieht sich auf die Art der Messung. Wenn die gewählten Fragen die relevanten Variablen nur eingeschränkt repräsentieren, werden auch die Modellprognosen unzuverlässig. Selbst bei objektiven Aufgaben kann ein Measurement-Bias durch unausgewogene Merkmalszuordnungen auftreten, wenn die erhobenen Daten systematisch vom Verständnis der Grundgesamtheit abweichen. Möglich ist, dass die vorgeschlagenen Bezeichnungen relevante Unterschiede zwischen gesellschaftlichen Gruppen ungenau erfassen oder dass kulturelle und individuelle Unterschiede systematische Abweichungen verursachen. So kann es beispielsweise vorkommen, dass die für einen Werbe-Claim generierten Texte den komplexen menschlichen Wahrnehmungen von Authentizität nicht gerecht werden, weil in den Trainingsdaten Authentizität nur durch eine einzige Frage und damit ungenau erhoben wurde. Solche Messfehler könnte man vermeiden, indem man mehrere konzeptionell verwandte Messmethoden im Sinne einer Triangulierung anwendet und die tatsächlichen Einschätzungen der Befragten besser versteht. Eine weitere Möglichkeit wäre die Prüfung der Varianz innerhalb klar definierter Zielgruppen.

- **Soziale Erwünschtheit (Social-Desirability-Bias)** Alle Trainingsdaten, die menschliche Wahrnehmungen, Meinungen oder historische Daten erfassen, sind anfällig für soziale Einflüsse. Biases treten auf, wenn die verfügbaren Daten bestehende Vorurteile, Normen oder Erwartungen im Sample widerspiegeln, was sich in unerwünschten Ergebnissen des KI-Modells niederschlägt. Es kam beispielsweise immer wieder zu verfälschten Ergebnissen, weil Algorithmen, die mit verzerrten Daten trainiert wurden, Frauen oder dunkelhäutige Personen bei Bankgeschäften, Bewerbungen oder Rechtsurteilen diskriminierten. Eine bewährte Methode zur Vermeidung dieses Bias besteht darin, kritische Merkmale wie Geschlecht oder ethnische Zugehörigkeit aus dem Modell herauszunehmen, um so Fairness und Gleichbehandlung zu gewährleisten. Diskriminierende Effekte sind allerdings trotzdem nicht auszuschließen, da kritische Merkmale mit weiteren Merkmalen des Modells korrelieren können. Abhilfe schaffen könnte eine Analyse der Korrelationsmatrix der zugrunde liegenden Trainingsdaten sowie eine Diskussion mit Experten über potenzielle Folgen der Anwendung von Algorithmen im realen Kontext.

Das wahre Potenzial generativer KI wird erst durch Feintuning der Modelle für einen spezifischen Markenkontext erschlossen.

Neben Biases in Basisdaten kann es auch bei speziell erhobenen Trainingsdaten zu Biases aufgrund sozialer (Un-)Erwünschtheit kommen. Befragte bewerten oft im Einklang mit kulturellen Normen, die Vorurteile oder Diskriminierung widerspiegeln. Wer sich bestimmter sozialer Erwartungen bewusst ist, könnte bei der Beantwortung von Fragen dazu neigen, diesen zu entsprechen. Am wahrscheinlichsten ist dies bei KI-Modellen, die sich mit Prognosen zu nachhaltigem, ethischem oder gesundheitsrelevantem Konsumentenverhalten beschäftigen. In diesen Fällen kann es vorkommen, dass Modelle übertriebene Präferenzen für Marketingkampagnen oder damit verbundene generierte Inhalt prognostizieren, z. B. bei Influencern, die Minderheiten repräsentieren, oder wenn es sich um ökologische und nachhaltige Produkte handelt.

› **Response-Bias** Während sich der Measurement-Bias auf Fragen und Antwortoptionen der Beurteilung bezieht, geht es beim Response-Bias um den Beurteilungsprozess an sich. Im Vergleich zu klassischen Befragungen ist die Erhebung von Trainingsdaten oft ein repetitiver und monotoner Prozess mit wenigen Fragen zu sich wiederholenden und variierenden Merkmalen. Bereits die Beurteilung objektiver Daten kann für menschliche Kodierer ermüdend und belastend sein, aber bei subjektiven Aufgaben ist die Messung noch deutlich komplexer und damit steigt auch die Wahrscheinlichkeit eines Response-Bias. Kodierer sind anfällig für bestimmte Antworttendenzen, wenn Fragebögen zu lang oder zu komplex sind. Typische Fälle sind „acquiescence" – die Neigung, Fragen unabhängig von deren Inhalt zuzustimmen – oder „disacquiescence", bei der Kodierer zu ablehnenden Antworten neigen oder gedankenlose und willkürliche Angaben machen.
Die Qualität der generierten KI-Ergebnisse hängt daher von der Menge und der Reihenfolge der Kodieraufgaben ab. Response-Bias kann die Effizienz und die Power eines Modells ernsthaft beeinträchtigen. Er ist besonders bedenklich, wenn das Modell nur durch wenige Antworten für einzelne Elemente, wie Videos, Bilder oder Texte, trainiert wurde, da GenAI eine ausreichende Varianz auf der Inhaltsebene erfordert.

Mehr Effektivität und Kosteneffizienz bei Modellen mit unverzerrten Daten Generative KI kann das Marketing nachhaltig verändern. Echte Wettbewerbsvorteile erreicht man, indem man Standard-GenAI-Modelle mit subjektiven Marketingmetriken auf den spezifischen Kontext der Marke trainiert. Eine entscheidende Voraussetzung ist jedoch, dass die Trainingsdaten von höchster Qualität sind. Auf verzerrte Trainingsdaten, die wiederum die Ergebnisse eines GenAI-Modells verfälschen können, sollte bei der Entwicklung von GenAI-Modellen ein besonderes Augenmerk gerichtet werden. Ein erstklassiges Modell ist dann erreichbar, wenn man potenzielle Biases bereits vor dem Training des Modells bewertet und eliminiert und die aktuelle Praxis der Fehleranalyse in der Post-Trainingsphase ergänzt. Durch ein vorgeschaltetes Bias-Screening wird das Training von KI-Modellen nicht nur effektiver, sondern auch kosteneffizienter.
Wir empfehlen den Einsatz interdisziplinärer Forschungsteams mit technischem Knowhow und Marktforschungskompetenzen sowie Kooperationen mit Softwareplattformen, um eine hochmoderne und reibungsfreie Datenannotation zu gewährleisten. Die vorgeschlagenen Maßnahmen tragen dazu bei, alle relevanten Perspektiven einzubinden und erfolgreiche GenAI-Anwendungsfälle zu entwickeln, die echte Wettbewerbsvorteile gegenüber Standardanwendungen bringen.

LITERATURHINWEISE

Feuerriegel, S., Hartmann, J., Janiesch, C. et al. (2023). Generative AI. Business & Information Systems Engineering, 66, 111–126. https://doi.org/10.1007/s12599-023-00834-7

Hartmann, J., Heitmann, M., Schamp, C., & Netzer, O. (2021). The power of brand selfies. Journal of Marketing Research, 58(6), 1159–1177.

Van Giffen, B., Herhausen, D., & Fahse, T. (2022). Overcoming the pitfalls and perils of algorithms: A classification of machine learning biases and mitigation methods. Journal of Business Research, 144, 93–106.