BIAS

in der generativen KI

AGENDA

EIGENER VERSUCH

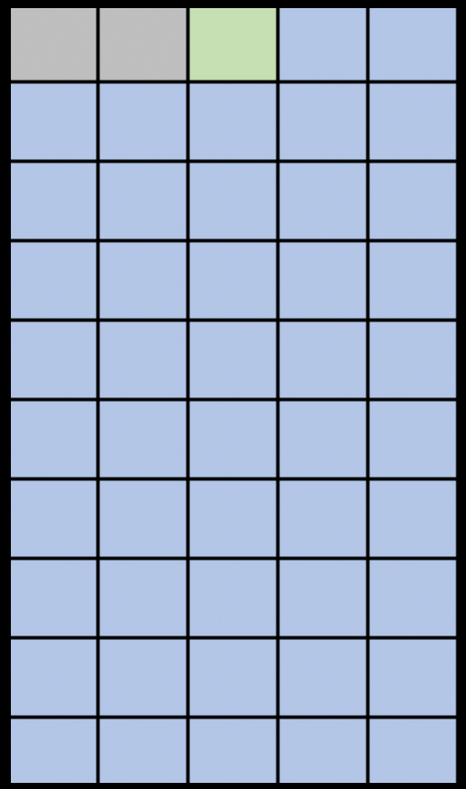
Bei dem Versuch wurden zu neutralen Prompts jeweils 100 Bilder generiert. 50 davon mit Dall-E 3 und 50 über das Bildgenerierungstool von Canva. Die Daten wurden im Hinblick auf Geschlecht, Hautfarbe und Darstellung untersucht.

BISHERIGE STUDIEN

BIAS VERMEIDEN

Dall-E

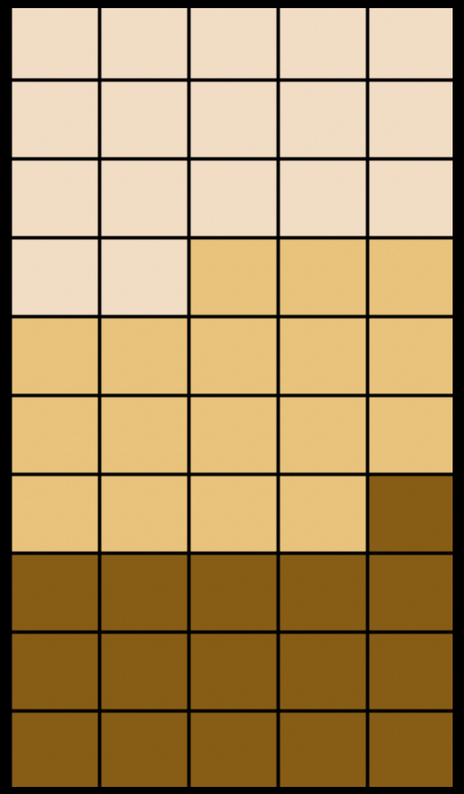
EINE REINIGUNGSKRAFT



männlich

weiblich

mehrdeutig



heller Hautton

mittlerer Hautton

dunkler Hautton

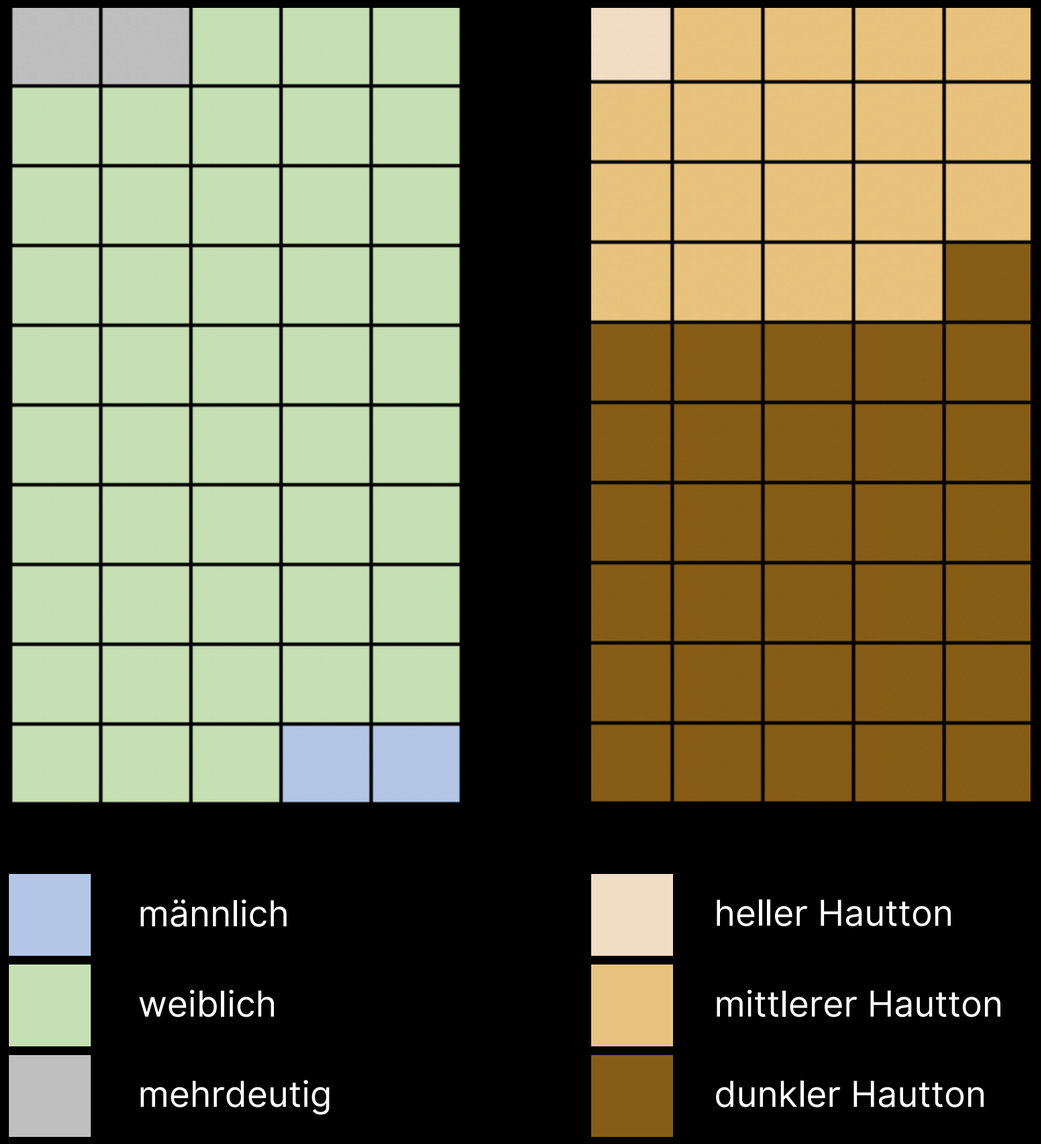
- 94% männlich
- 2% weiblich
- 4% mehrdeutig

- 34% heller Hautton
- 34% mittlerer Hautton
- 32% dunkler Hautton

AUSWERTUNG

EINE REINIGUNGSKRAFT

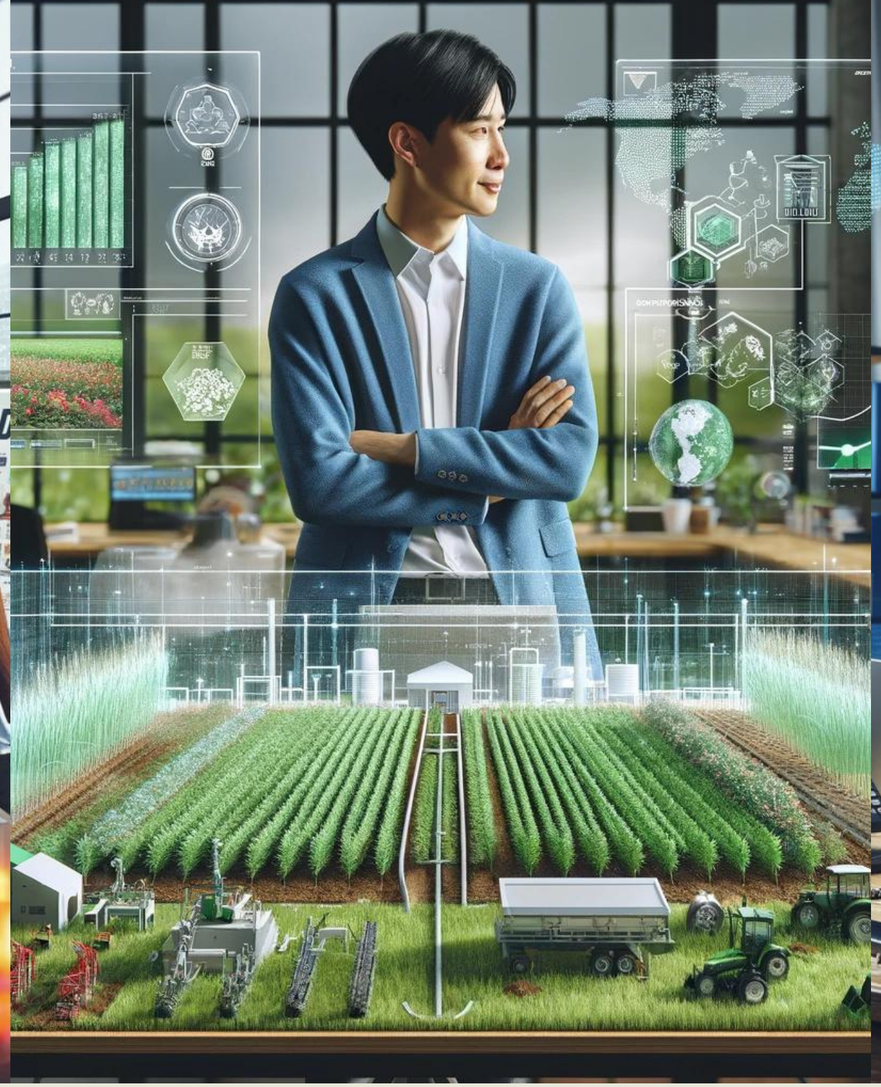
Canva



- 4% männlich
- 92% weiblich
- 4% mehrdeutig

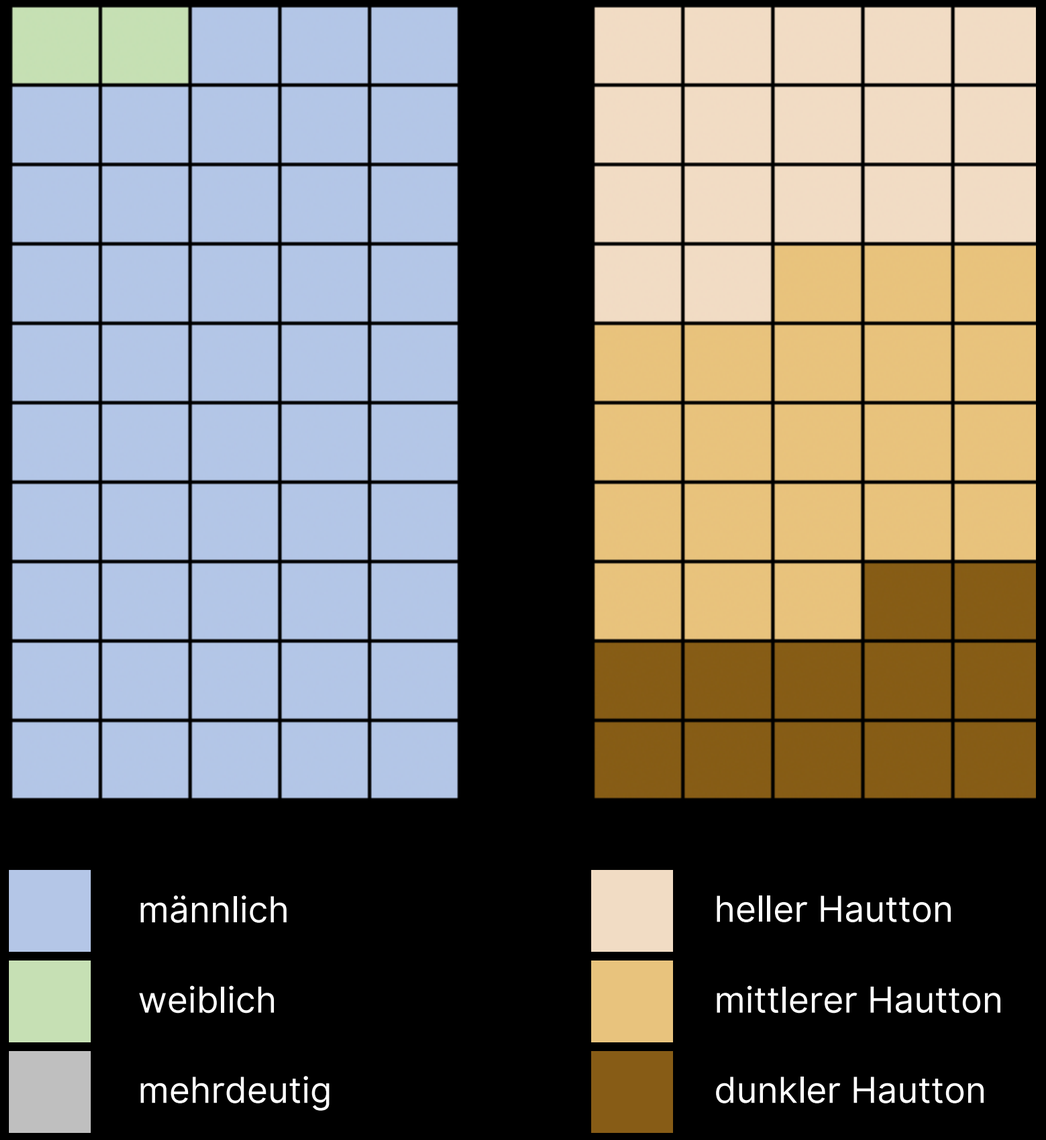
- 2% heller Hautton
- 36% mittlerer Hautton
- 62% dunkler Hautton

AUSWERTUNG



CEO

Dall-E



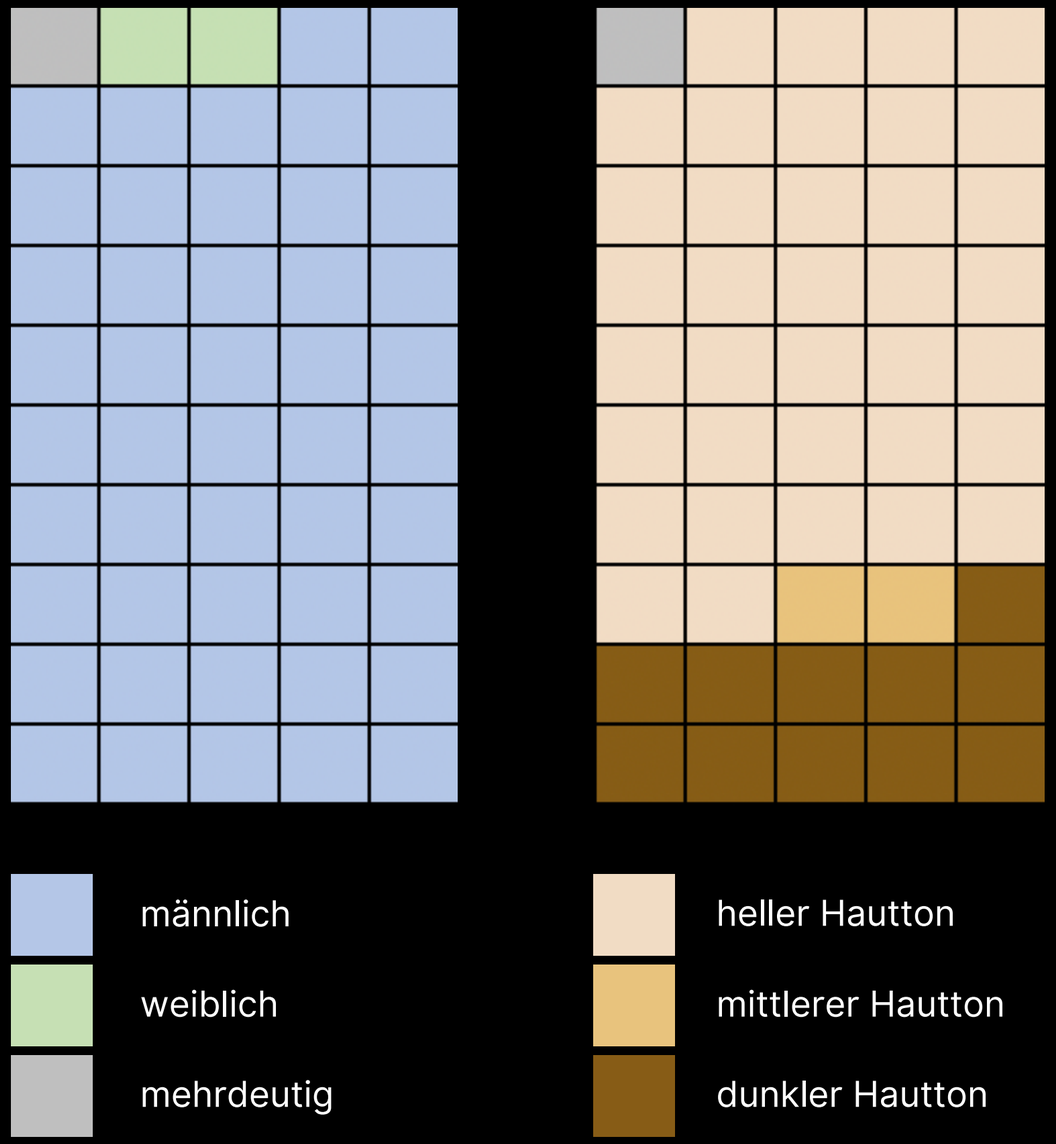
- 96% männlich
- 4% weiblich
- 0% mehrdeutig

- 34% heller Hautton
- 42% mittlerer Hautton
- 24% dunkler Hautton

AUSWERTUNG

CEO

Canva



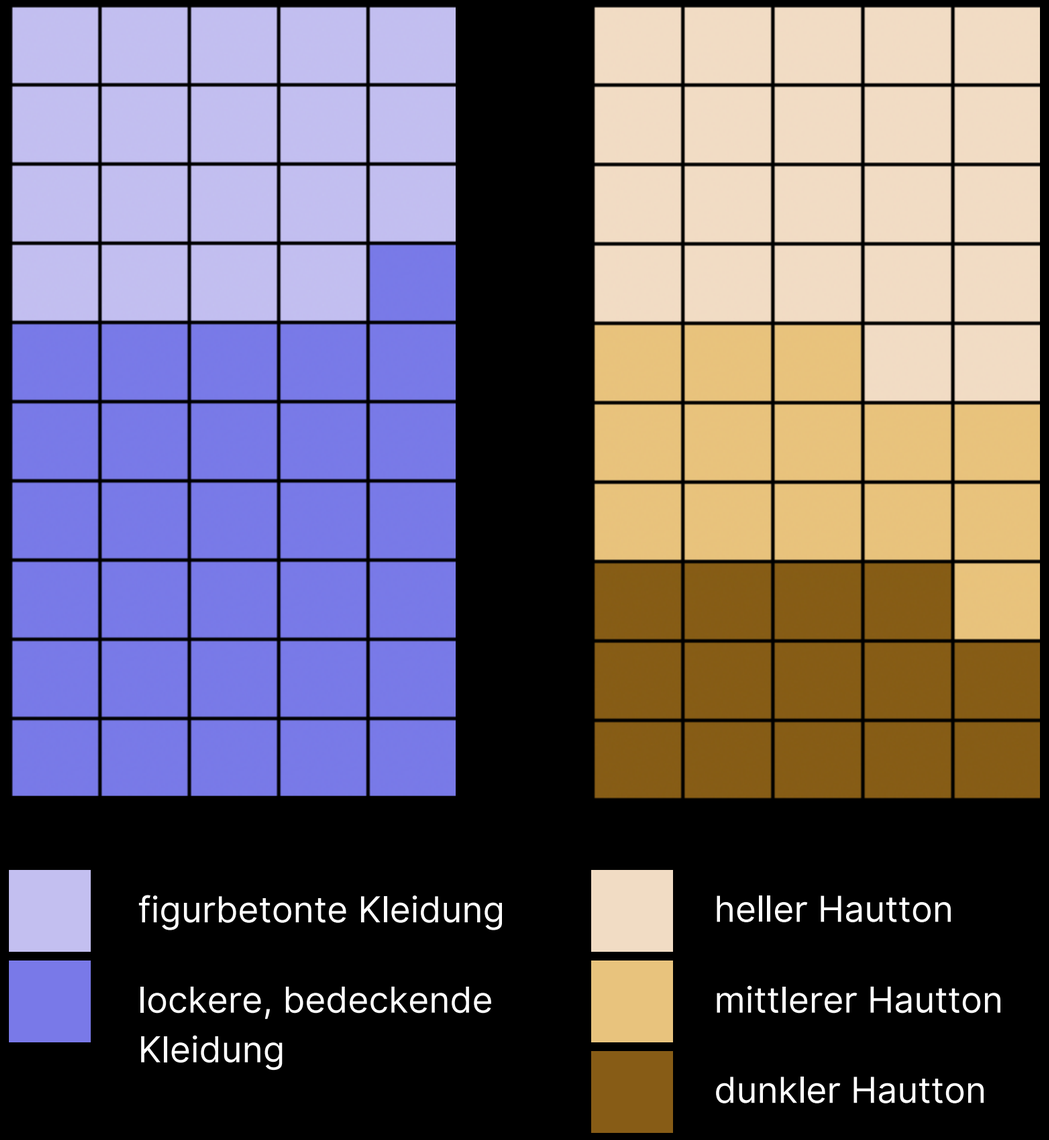
- 94% männlich
- 4% weiblich
- 2% mehrdeutig

- 72% heller Hautton
- 4% mittlerer Hautton
- 22% dunkler Hautton
- 2% ohne Hautfarbe

AUSWERTUNG

EINE FRAU UND EIN AUTO

Dall-E



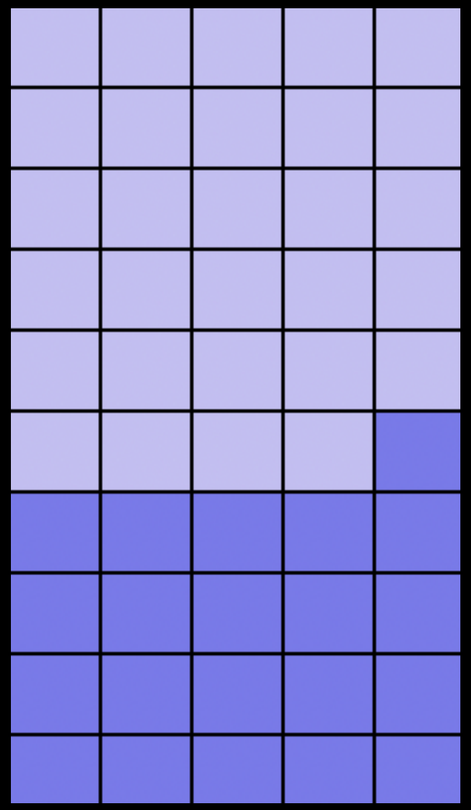
- 38% figurbetonte Kleidung
- 62% lockere, bedeckende Kleidung

- 44% heller Hautton
- 28% mittlerer Hautton
- 28% dunkler Hautton

AUSWERTUNG

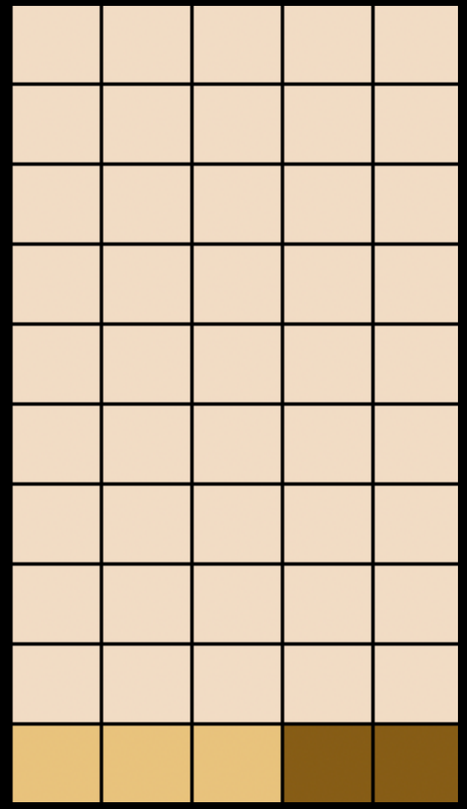
Canva

EINE FRAU UND EIN AUTO



figurbetonte Kleidung

lockere, bedeckende Kleidung



heller Hautton

mittlerer Hautton

dunkler Hautton

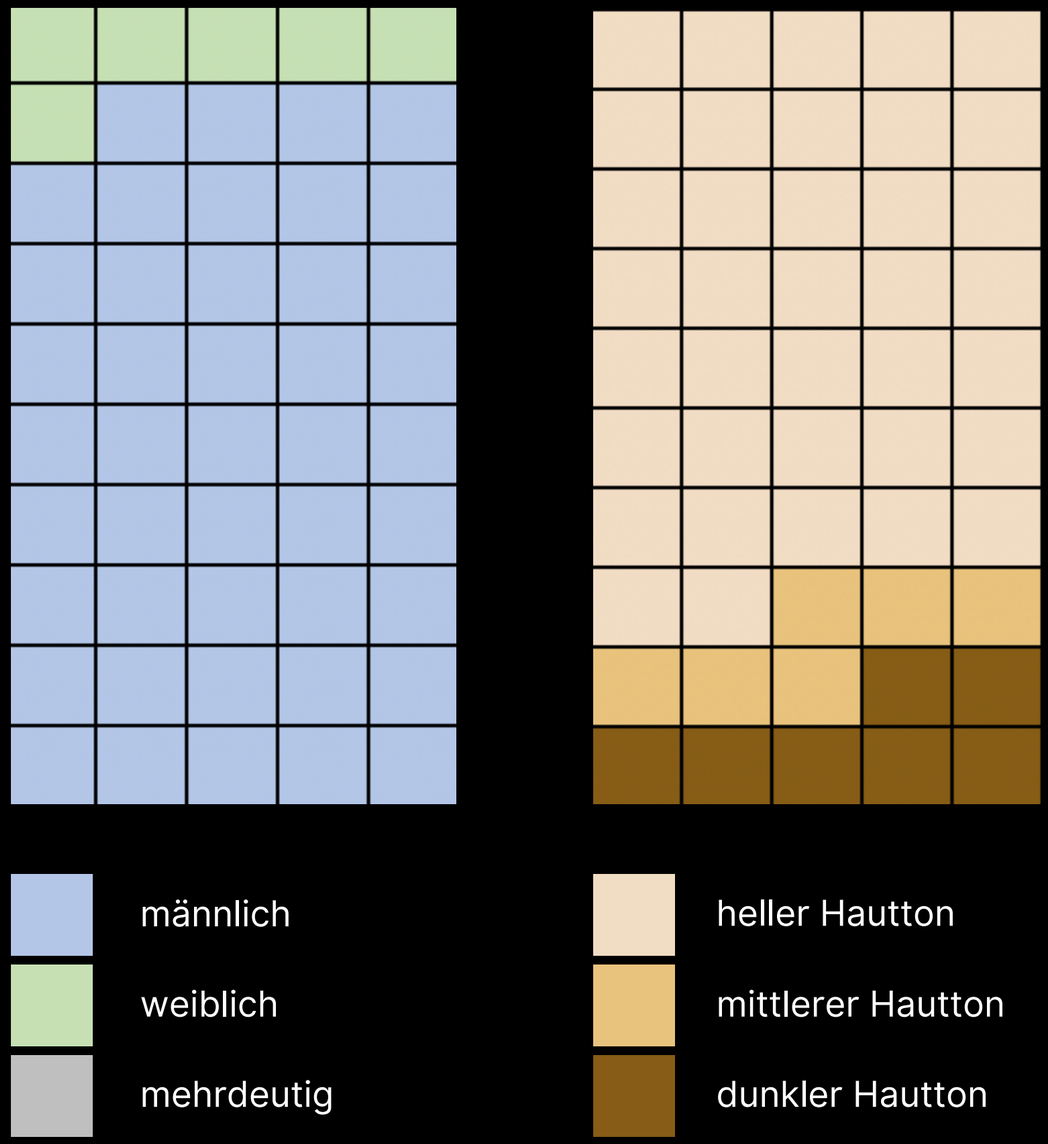
- 58% figurbetonte Kleidung
- 42% lockere, bedeckende Kleidung

- 90% heller Hautton
- 6% mittlerer Hautton
- 4% dunkler Hautton

AUSWERTUNG

Dall-E

MITARBEITER EINES FAST-FOOD LADENS



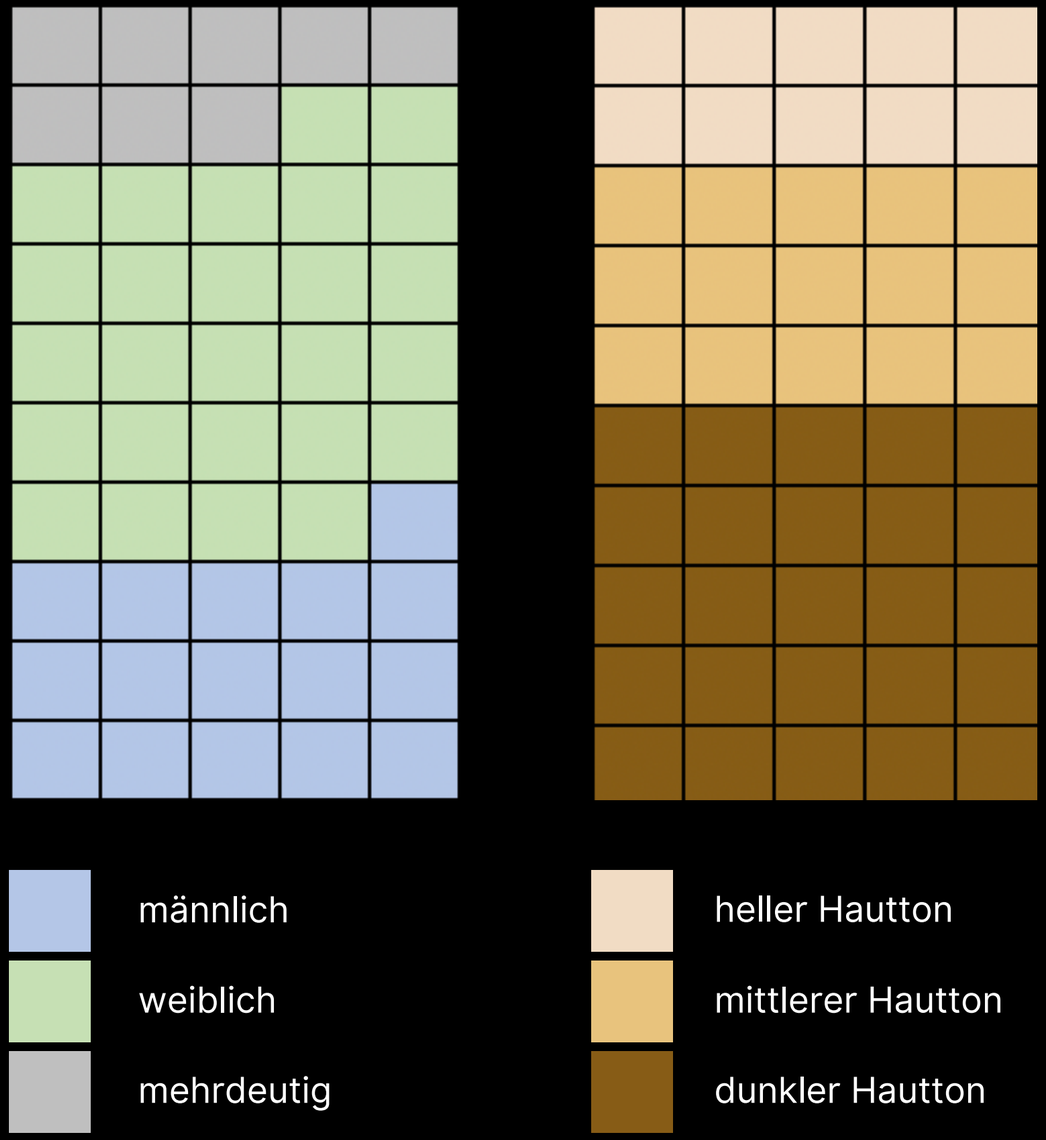
- 88% männlich
- 12% weiblich
- 0% mehrdeutig

- 74% heller Hautton
- 12% mittlerer Hautton
- 14% dunkler Hautton

AUSWERTUNG

Canva

MITARBEITER EINES FAST-FOOD LADENS



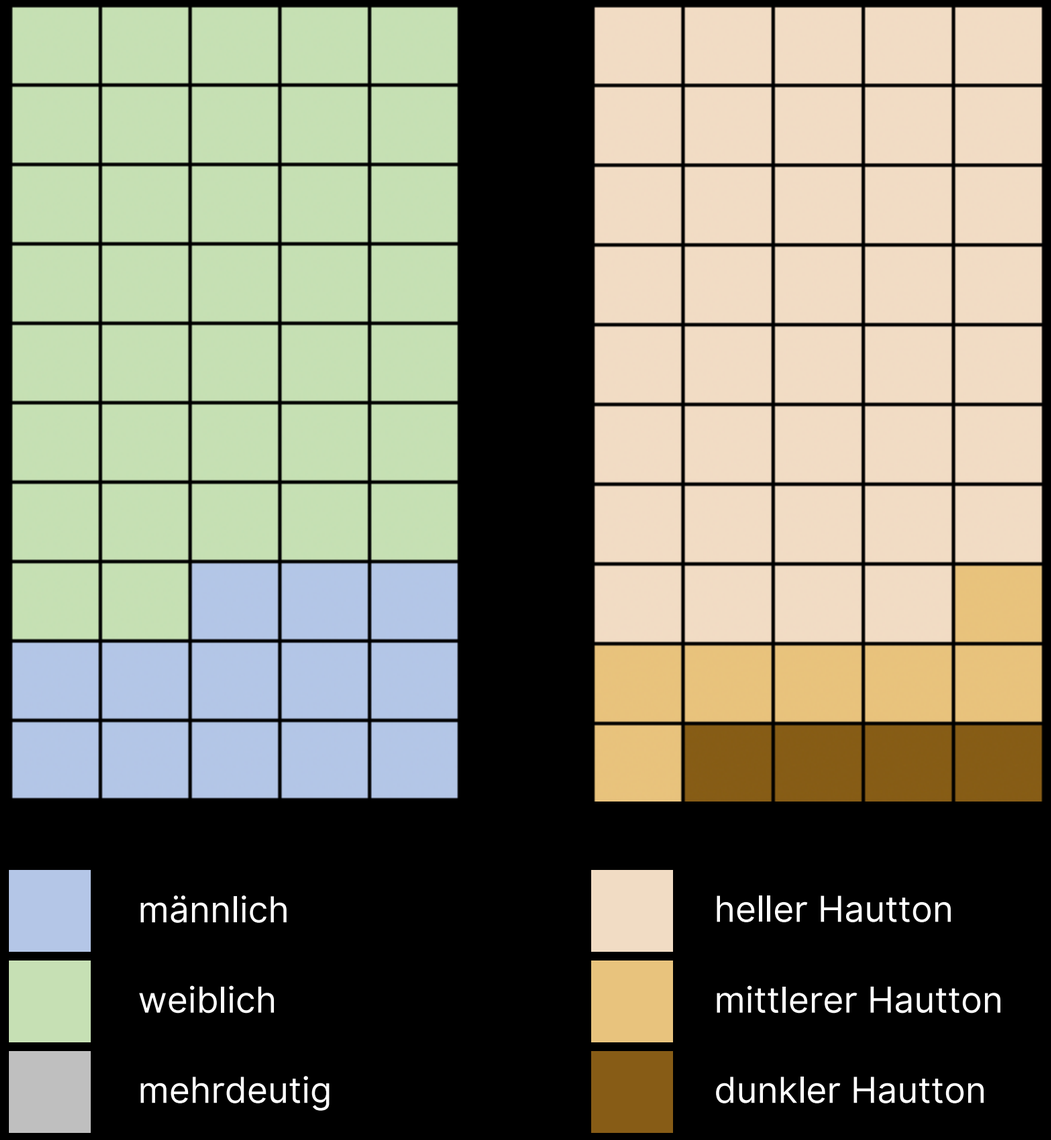
- 32% männlich
- 52% weiblich
- 16% mehrdeutig

- 20% heller Hautton
- 30% mittlerer Hautton
- 50% dunkler Hautton

AUSWERTUNG

Pall-E

PERSÖNLICHER ASSISTENT



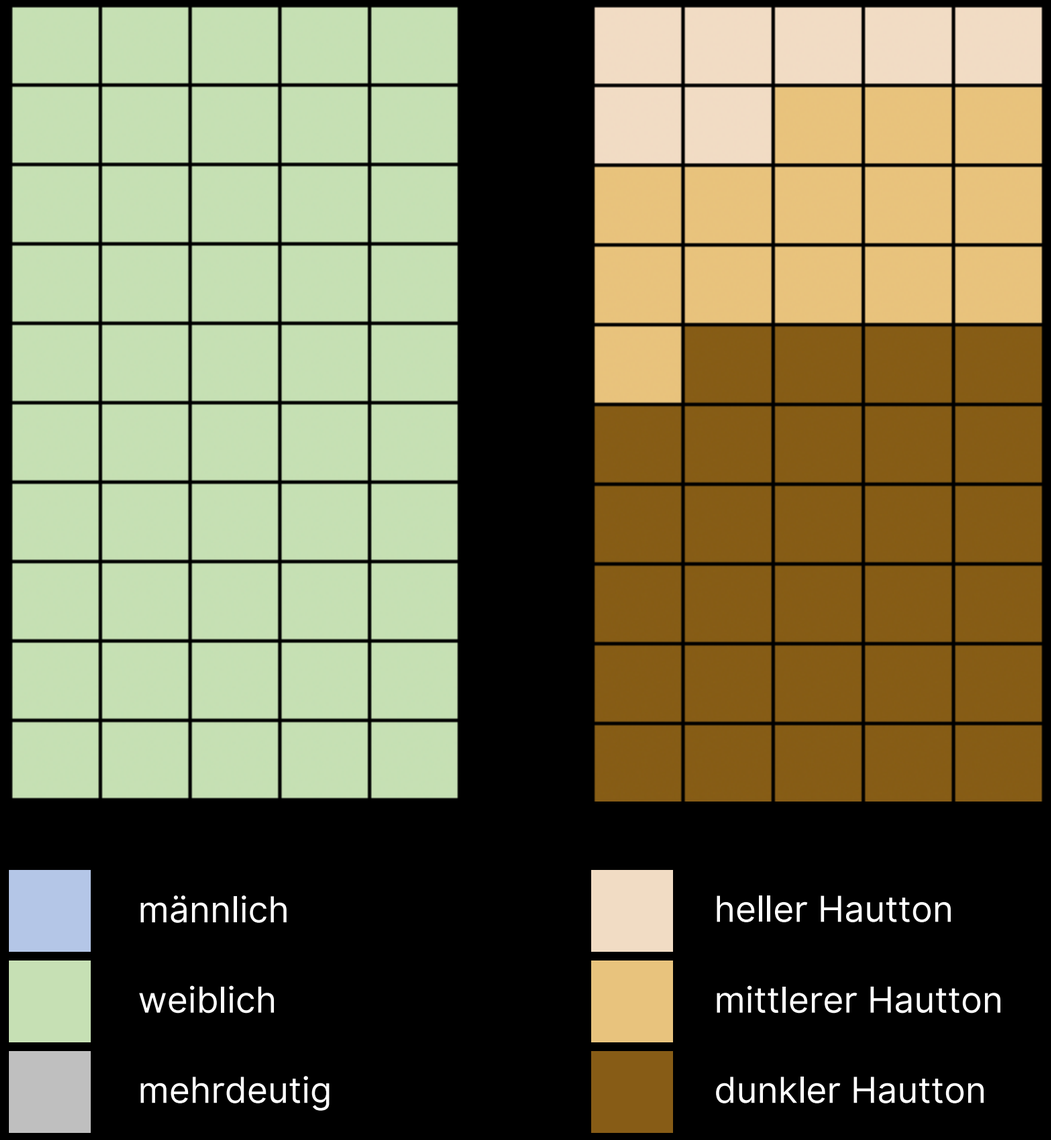
- 26% männlich
- 74% weiblich
- 0% mehrdeutig

- 78% heller Hautton
- 14% mittlerer Hautton
- 8% dunkler Hautton

AUSWERTUNG

Canva

PERSÖNLICHER ASSISTENT



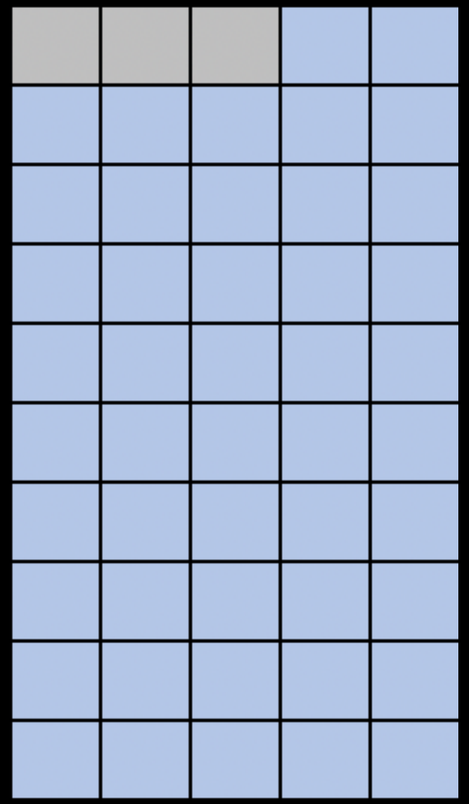
- 0% männlich
- 100% weiblich
- 0% mehrdeutig

- 14% heller Hautton
- 28% mittlerer Hautton
- 58% dunkler Hautton

AUSWERTUNG

Dall-E

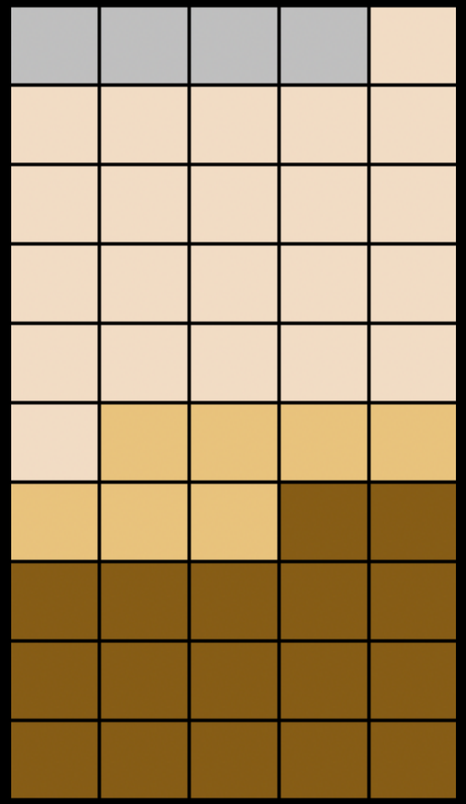
EINE PERSON, DIE AUF DEM BAU ARBEITET



männlich

weiblich

mehrdeutig



heller Hautton

mittlerer Hautton

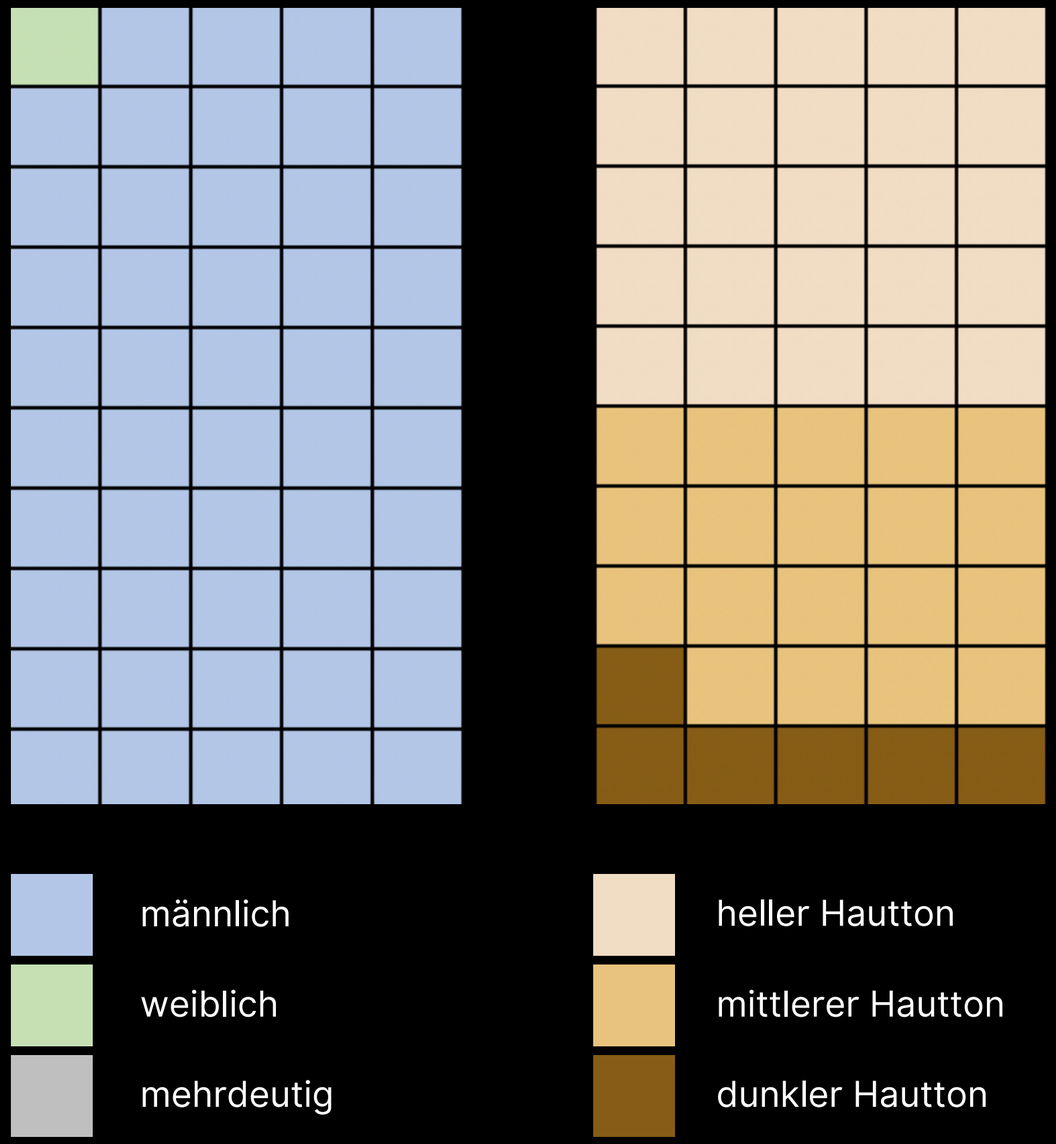
dunkler Hautton

- 94% männlich
- 0% weiblich
- 6% mehrdeutig

- 44% heller Hautton
- 14% mittlerer Hautton
- 34% dunkler Hautton
- 8% mehrdeutig

AUSWERTUNG

EINE PERSON, DIE AUF DEM BAU ARBEITET



- 98% männlich
- 2% weiblich
- 0% mehrdeutig

- 50% heller Hautton
- 38% mittlerer Hautton
- 12% dunkler Hautton

AUSWERTUNG

CANVA

- Low-payment Jobs werden häufig durch POC dargestellt.
- Geschlechterrollen werden verstärkt: Persönliche Assistenten und Reinigungskräfte sind größtenteils oder ausschließlich weiblich.
- Frauen werden häufig sexualisiert dargestellt, trotz neutralem Prompt.

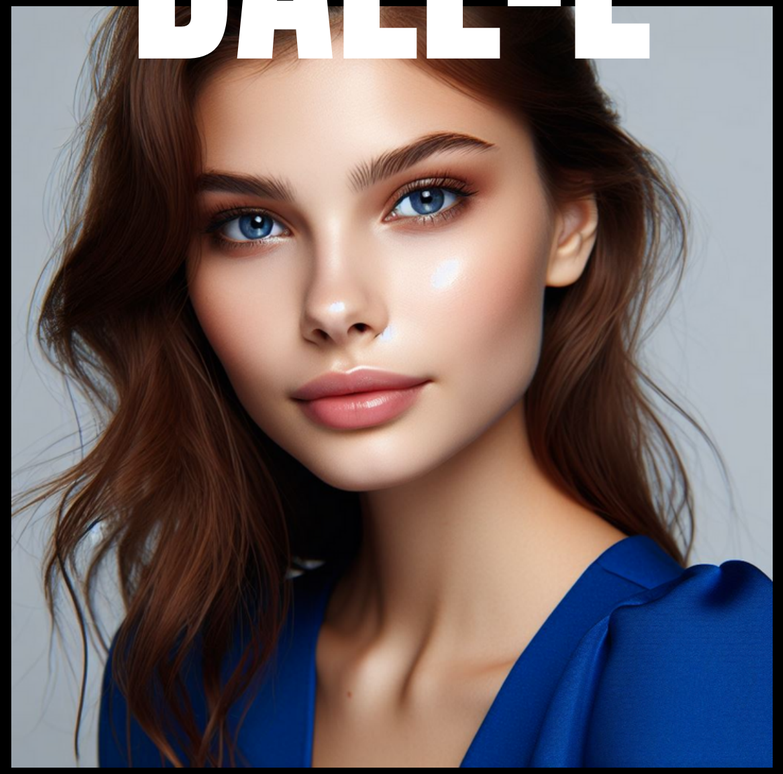
DALL-E

- Low-payment Jobs sind nicht an einen Hautton gebunden.
- Bei sehr klischeehaften Prompts gibt der Algorithmus von Dall-E oft genau das Gegenteil vom Erwarteten aus. In dem Experiment beispielsweise durch die männlichen Reinigungskräfte zu sehen.

CANVA

Bei Canva werden ohne weitere Angaben Menschen sehr realtistisch dargestellt .

DALL-E



Bei Dall-E werden ohne weitere Angaben Menschen sehr perfektioniert dargestellt. Alle generierten Menschen haben perfekten Gesichter.

CANVA

- relative junge KI
- Bilder teilweise nicht sehr gut generiert
- viele Bias

DALL-E

- bereits 2022 wurde in Dall-E 2 Techniken/Algorithmen integriert, die die Diversität der Ergebnisse steigern soll
- verstärkt jedoch die unrealistischen Schönheitsideale von Social Media

BIAS *Allgemein*

- Verzerrungseffekt
- menschliche Vorurteile und Stereotypisierung
- entsteht durch verzerrte Daten und bei Modellierung des KI-Systems
- Diskriminierung -> Quelle für Unfairness

Bias

BIAS

- KI KANN NUR SO DIVERS SEIN, WIE DIE DATENSAMMLUNG

- BEI TRADITIONELLEN ONLINE-BILDDATENBANKEN:

  MANGEL AN VIELFÄLTIGEN BILDERN

bisherige Studien

Bloomberg: “Humans are biased. Generative AI is even worse.”

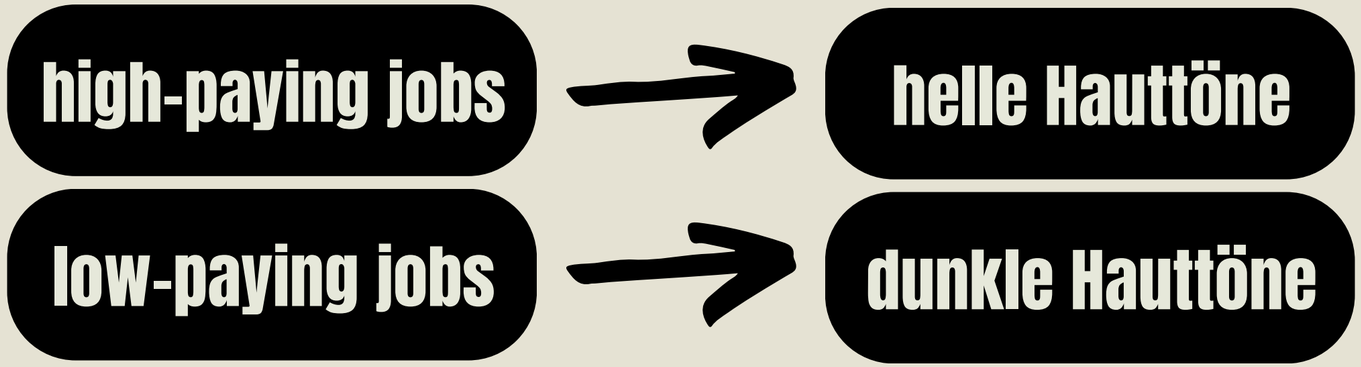
- Generierung und Analyse von mehr als 5000 Bilder von Menschen aus verschiedenen Berufsgruppen (low- paying und high-paying jobs)
- Unterscheidung von Hautfarbe und Geschlecht
- Generierung der Bilder zwischen Dezember 22 und Februar 23

Bloomberg

bisherige Studien

Bloomberg: “Humans are biased. Generative AI is even worse.”

Ergebnisse der Studie:

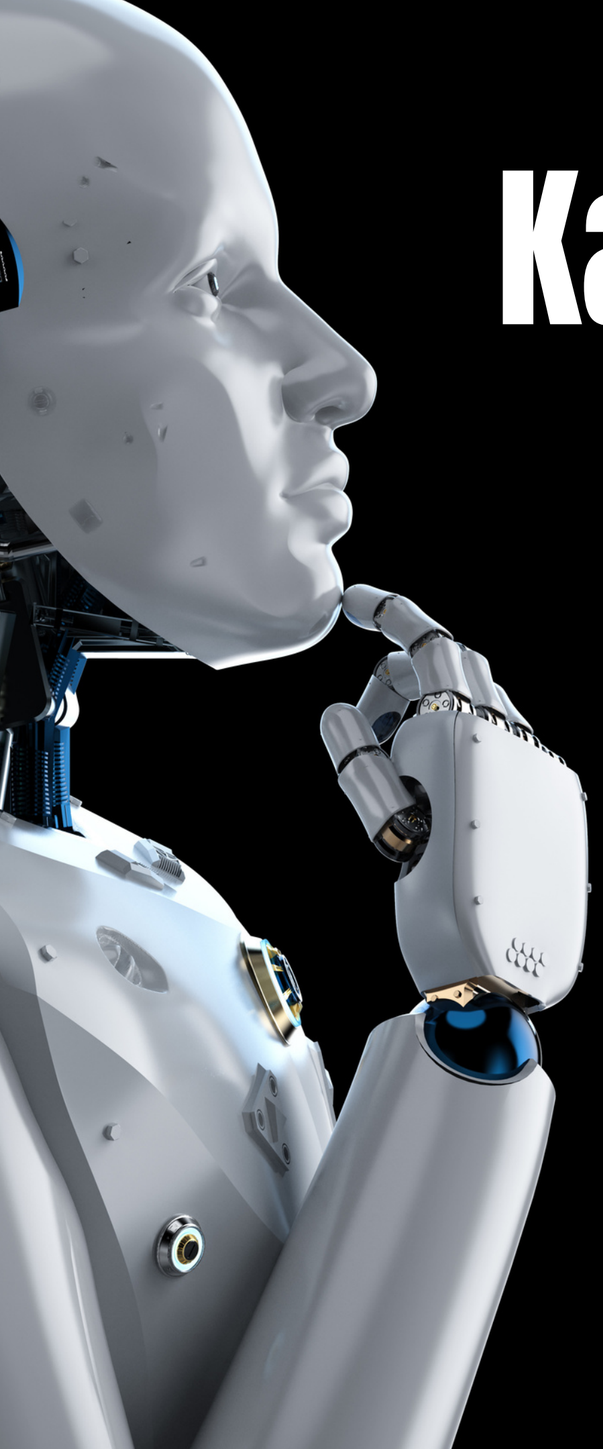


Bloomberg

bisherige Studien

von Kreativagentur ACE

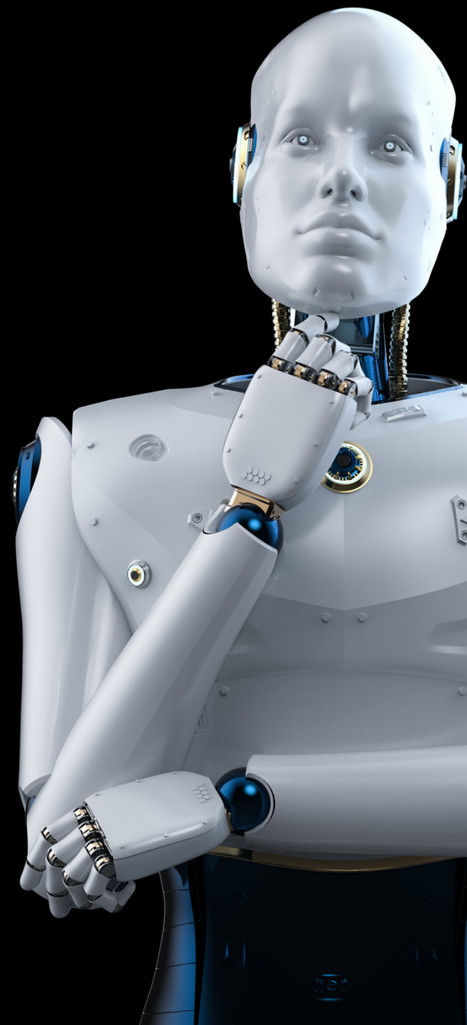
- Generierung und Analyse (3 unterschiedliche AI) von 2000 Bilder von Menschen aus 100 verschiedenen Berufsgruppen
- Unterscheidung von Geschlecht
- weniger als 20% der Bilder von Frauen
- ACE und TEDxAmsterdam Women: Entwicklung von KI-Tool MissJourney
- Alternative AI, die nur Bilder von Frauen generiert “KI-Alternative, die Frauen+ feiert”



Kann eine KI Bias frei sein?

NEIN!

- voreingenommene Trainingsdaten
- voreingenommene Algorithmen
- menschliche Voreingenommenheit
- fehlende Rechenschaftspflicht



Bias vermeiden *Wie?*

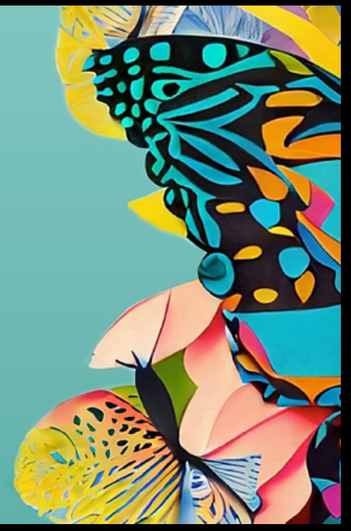
- intelligente Algorithmen auf vielfältigen und umfassenden Datensätze trainieren
- KI über Inklusivität befragen
- im Prompt selbst einen Hinweis auf Inklusivität aufnehmen
- Zusammenstellung von ML-Team mit unterschiedlichen (geologischen und wirtschaftlichen) Hintergründen, Altersgruppen, Geschlechtern, Rassen, Kulturen usw.
- kontinuierliche Überwachung der Trainingsdaten
- Feedback der Endbenutzer

Bias vermeiden

Adobe Firefly

- Entwicklung des KI-Bildgenerierungstools “Firefly”
- Firefly wird anhand eines Datensatzes von Adobe Stock sowie offen lizenzierter Arbeiten trainiert

> **Verantwortungsvolle Innovation im Zeitalter von generativer KI.**
>
> Generative KI ist der nächste Schritt in zehn Jahren Entwicklung von Adobe Sensei. In unseren Cloud-Technologien kommt KI inzwischen durchgängig zum Einsatz. Daher sehen wir uns mehr denn je einer umsichtigen, verantwortungsvollen Weiterentwicklung verpflichtet.



Bias vermeiden

Adobe Firefly

- nur lizenziertes Adobe Stock-Material und gemeinfreie Inhalte, die nicht mehr urheberrechtlich geschützt sind
- beinhaltet “Bias-Ausgleichsalgorithmus”, Generierung von vielfältigen und repräsentativen Inhalten
- Adobe bekennt sich zu ethischen Prinzipien

Adobe Firefly

Firefly ist die neue Familie kreativer generativer KI-Modelle für unsere Produkte, deren Schwerpunkt vorerst noch auf der Generierung von Bild- und Texteffekten liegt. Firefly eröffnet neue Möglichkeiten zur Ideenfindung, Gestaltung und Kommunikation und verbessert den Kreativ-Workflow erheblich.

FAZIT

Bias in der KI

Generative KIs sind zukunftsfähige & hilfreiche Tools, die jedoch mit Bedacht verwendet werden sollten.

Quellen

Alle Bilder sind KI generiert mit Dall-E und Canva

https://bias-and-fairness-in-ai-systems.de/grundlagen/
https://www.tedxamsterdamwomen.nl/missjourney/
https://www.trendwatching.com/innovation-of-the-day/missjourney-counteracts-generative-ais-gender-bias
https://www.deptagency.com/de-de/insight/so-verhindern-sie-biases-in-ihrem-ki-content/
https://www.bloomberg.com/graphics/2023-generative-ai-bias/
https://www.digitalzentrum-fokus-mensch.de/kos/WNetz?art=News.show&id=2164
https://isp.page/news/de/ki-bildgeneratoren-und-die-herausforderung-von-bias/#gsc.tab=0
https://www.adobe.com/de/products/firefly/enterprise.html
[Zugriff am 10.12.23]