Bias of large language models: Evaluation of the desert-based approach to fairness

Presented at LLM and Philosophy Conference,
Kanazawa, Japan
September 2024

Chong Ho Alex Yu, Ph.D., D. Phil.

HAWAIʻI PACIFIC UNIVERSITY

Bias and fairness: Multifaceted concepts

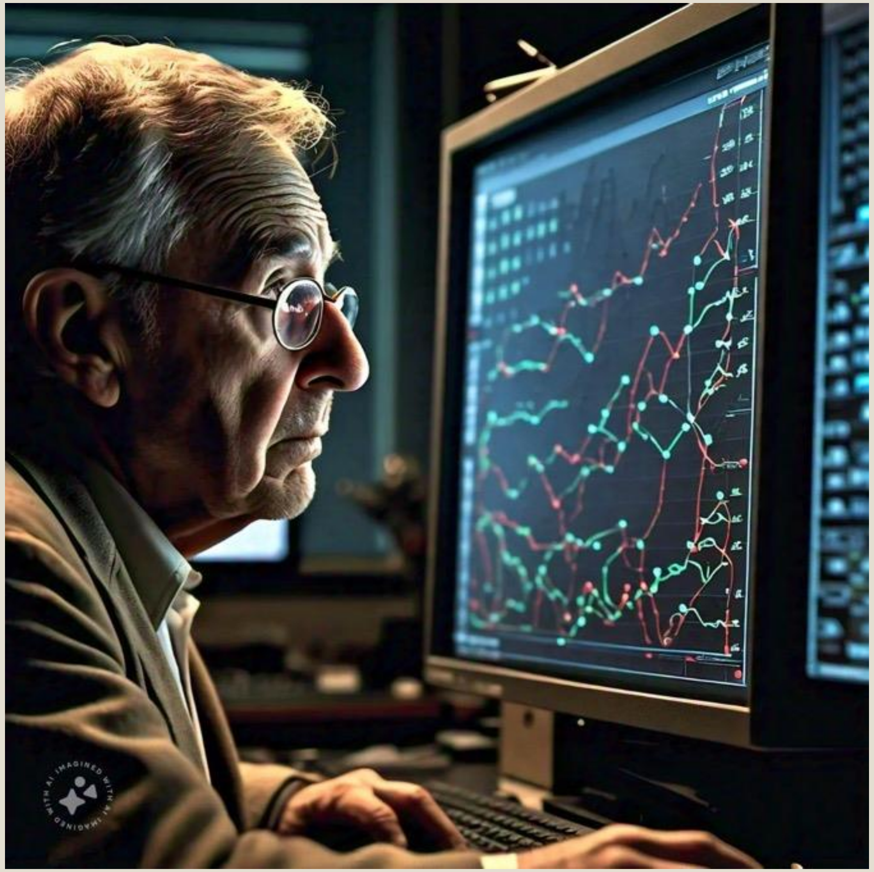
HAWAIʻI PACIFIC UNIVERSITY

Bias and Fairness: See-saw relationship

- Bias refers to systematic errors or prejudices that result in unfair outcomes.
- Fairness is a multifaceted concept that can be interpreted in various ways depending on the context and stakeholders involved.

Bias and fairness are in a see-saw relationship: As bias increases, fairness decreases, and vice versa.

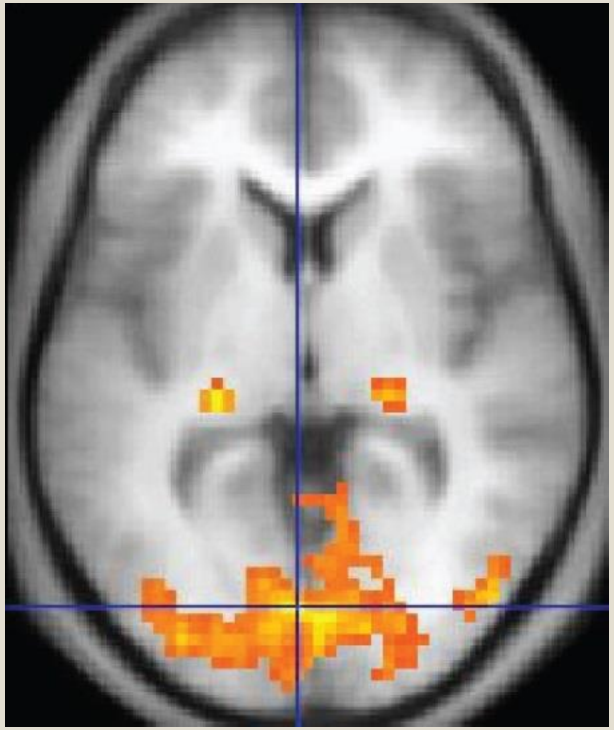
Bias



- **Sample Bias:**
  - Definition: Occurs when the data used to train a model is not representative of the population it is meant to serve. This leads to skewed results and decisions.
  - Example: If an AI system for job recruitment is trained predominantly on data from male candidates, it may unfairly favor male applicants over female applicants.

Bias

- **Measurement Bias:**
  - Definition: Arises when there are inaccuracies in the data collection process, leading to errors in the data that can affect model outcomes.
  - Example 1: Inaccurate labeling of medical images in a dataset can result in an AI system that misdiagnoses conditions.
  - Example 2: Measure people’s intelligence using a particular IQ test, but the test is not adequately validated, and its psychometric properties are questionable.



Bias

- **Algorithmic Bias:**
  - Definition: Occurs when the algorithm itself contributes to biased outcomes, often due to design choices, assumptions, or the way the algorithm processes data.
  - Example: An AI system used for criminal sentencing that relies on historical data may perpetuate existing biases in the justice system, leading to harsher sentences for certain demographic groups.

Bias

- **Label Bias:**
  - Definition: Occurs when the labels in the training data are biased, which can affect the performance and fairness of the model.
  - Example: If human annotators label data based on their subjective judgments or societal biases, the AI model will learn and perpetuate these biases.
- These issues are not unique to AI and data science; they also arise in classical statistics, which grapples with challenges such as sampling errors and measurement errors.

Fairness

- **Distributive Fairness**:
  - Ensure that the outputs or decisions made by AI models are equitable across different demographic groups, such as race, gender, age, and socio-economic status.
  - Example: An AI system used for loan approvals should not disproportionately deny loans to applicants from certain racial or ethnic groups, given that those applicants are as qualified as others.
  - Fairness is about **equal opportunities**, not equal outcomes. In this sense demographic parity is not guarantied.

Fairness

- **Procedural Fairness:**
  - This definition emphasizes the processes and procedures leading to decisions made by AI systems. Procedural fairness ensures that the methods used to collect data, train models, and make decisions are transparent, consistent, and unbiased.
  - Example: Ensuring that the training data for an AI hiring tool is diverse and representative of the population it will serve.
  - In some cases, even if the outcome is just, one cannot override the procedure. For example, a police officer cannot break into the home of a criminal without a search warrant.

Fairness

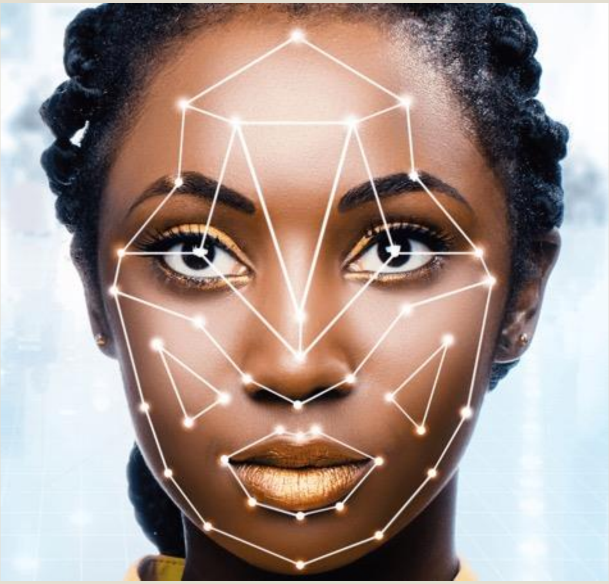
- **Individual Fairness:**
  - This concept focuses on treating similar individuals similarly. It requires that an AI system's decisions should be consistent for individuals who are alike in relevant aspects.
  - Example: Two job applicants with similar qualifications and experience should have the same chances of being selected by an AI-based recruitment system.

Fairness

- **Individual Fairness:**
  - The challenge lies in defining what constitutes "similarity" between individuals and the weighing scores.
  - Example 1: In a university admission system, should a student with a 4.0 GPA but few extracurriculars be considered "similar" to a student with a 3.7 GPA but extensive leadership experience?
  - Example 2: In a loan application system, is someone with a high income but low credit score "similar" to someone with average income and excellent credit?
  - When one criterion carries more weight than others, the outcome may be viewed as unfair.

Fairness

- **Fairness in Representation:**
  - Ensure that the data used to train AI systems is representative of the diverse populations that the system will serve. It addresses biases that may arise from underrepresentation or overrepresentation of certain groups.
  - Example: Ensuring that a facial recognition system is trained on a diverse dataset that includes images of people from various ethnic backgrounds to avoid biases in recognition accuracy.



Fairness

- **Counterfactual Fairness:**
  - This definition focuses on ensuring that an individual would receive the same decision in a counterfactual world where a sensitive attribute (like race or gender) is different.
  - Example: An AI system determining credit scores should produce the same score for an individual if their race were hypothetically changed.

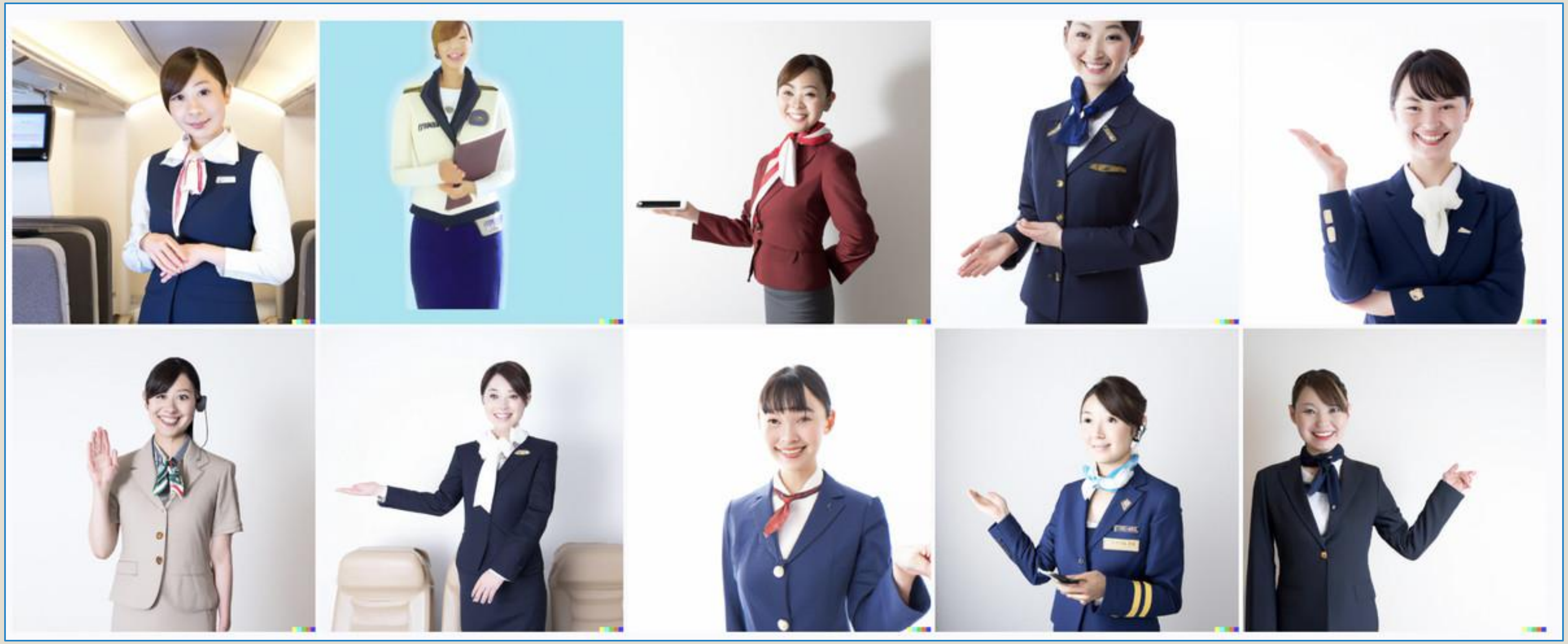
Bias against females in generative AI art tools

Generative AI art tools and LLM

- Most people associate LLMs with chatbots, such as ChatGPT, Claude, Google Gemini, and Perplexity.
- Generative AI art tools, such as DALL.E, Midjoruney, Stable Diffusion, and Adobe Firefly, also utilize LLMs to process prompts.
- In OpenAI, ChatGPT and DALL ·E share a common foundational approach in using large language models (LLMs) to process and understand natural language inputs. Both systems are trained on vast amounts of textual data to learn the structure, meaning, and relationships within language.

Bias towards attractive women

- In the earlier version of DALLE, when you requested a picture of a flight attendant, the system generated images of beautiful young Asian females.

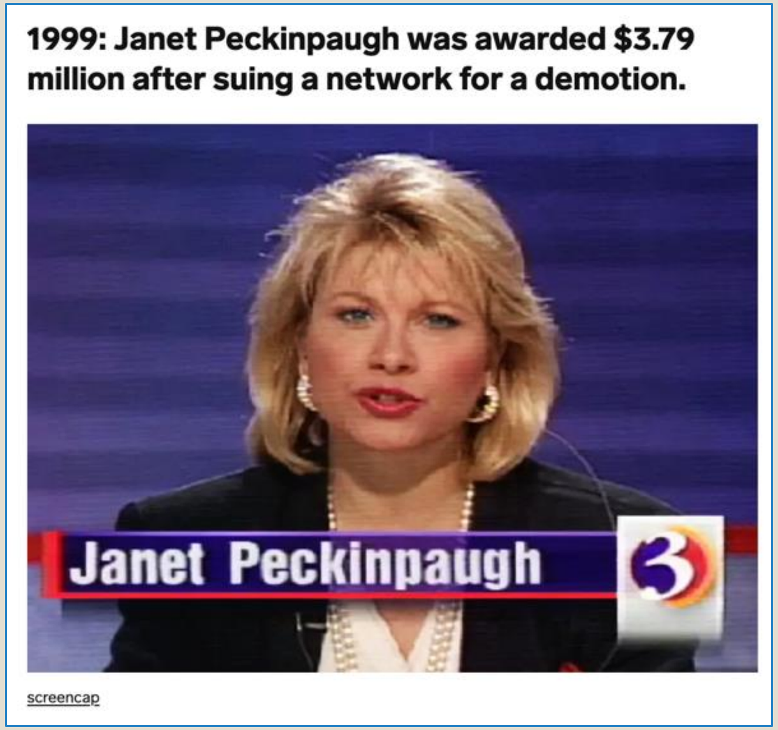


Bias towards attractive women

- This is not a unique issue in AI.
- Christine Craft was a TV host, but in August 1981, a market research team decided she was too old and unattractive, leading to her being fired from her anchor position. Craft filed a lawsuit against the network.
- A harsh reality: Audiences often expect to see young and attractive women on TV.
- Nonetheless, today the appearance of anchors on American television networks has become much more diverse.

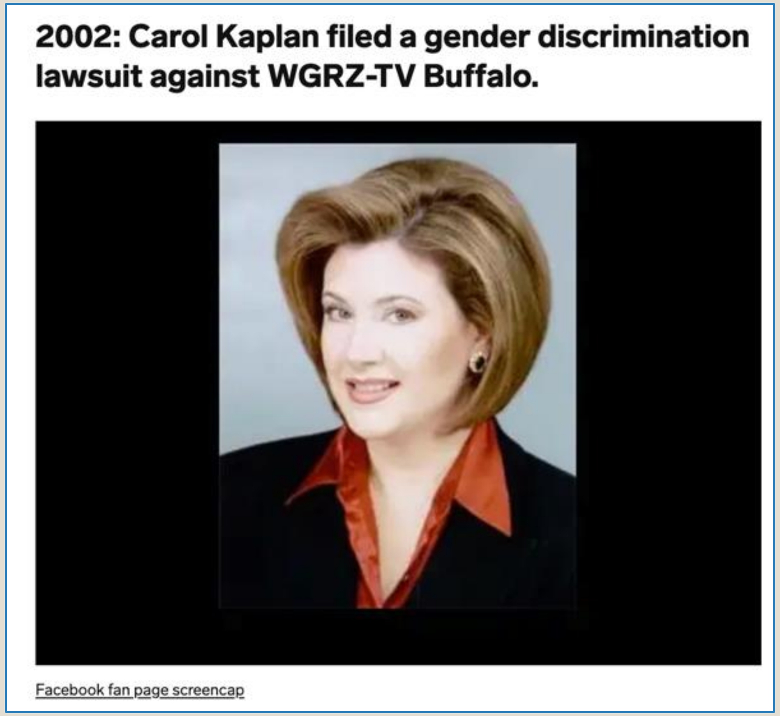
Bias towards attractive women

- Similar cases happened many times.



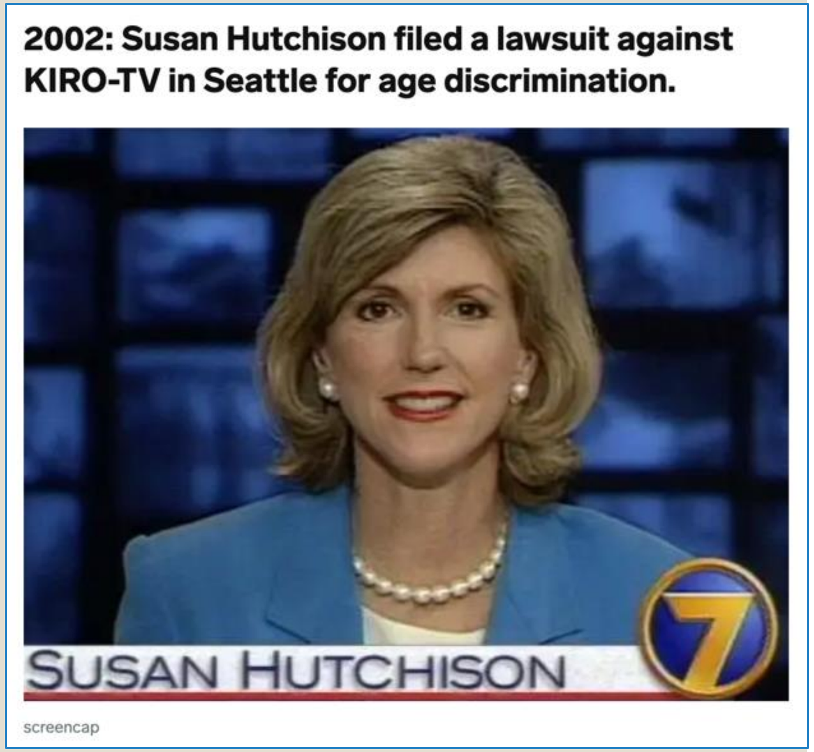
1999: Janet Peckinpaugh was awarded $3.79 million after suing a network for a demotion.

screencap



2002: Carol Kaplan filed a gender discrimination lawsuit against WGRZ-TV Buffalo.

Facebook fan page screencap



2002: Susan Hutchison filed a lawsuit against KIRO-TV in Seattle for age discrimination.

screencap

Bias towards attractive women

- In the past, American Airlines prioritized appearance as the most important factor in recruiting flight attendants, using the beautiful image of women to attract passengers.
- As a result, flight attendants' uniforms were usually tight-fitting, and they were required to wear high heels.
- Flight attendants faced termination if they were in their thirties, not petite enough, or planning to get married.

Bias towards attractive women

- In the 1960s, as Americans' awareness of equal rights gradually increased, these appearance requirements were abolished, and now US flight attendants are very diverse.
- Asian airlines still retains the tradition of hiring young and beautiful women as flight attendants.
- Argument: DALL.E simply reflects the reality.

Desert-based approach

- According to the desert-based approach to fairness and bias, individuals should receive deserved outcomes based on their features. For instance, a diligent student might deserve a high grade for submitting an excellent paper.
- Similarly, from a marketing perspective, attributes like youthfulness and attractiveness are sought after by consumers, leading to the proliferation of such images on the internet.

Evolutionary aesthetics

- Some argue that depicting images of young and attractive women is an objectification of women, and thus it is biased and unethical.
- However, it simply reflects our **natural psychological tendencies**. When watching a movie, we expect to see attractive stars; when traveling, we seek beautiful landscapes; and when visiting a museum, we desire to view beautiful artworks.
- Evolutionary aesthetics**:** This field, based on evolutionary psychology, studies the evolutionary origins of aesthetic preferences. Some researchers argue that our attraction to certain physical features may have evolutionary roots related to health and reproductive fitness.

Some objective standards?

- Beauty is a **standard merit** that deserves higher priority. There is no absolute standard for beauty. While we cannot determine whether Jodie Comer, Brigitte Bardot, or Sayuri Yoshinaga is more beautiful, we can differentiate a pretty woman and an odd-looking one.
- Some argue that beauty and attractiveness are objective qualities that can be measured and quantified. If so, AI systems can identify and reproduce those qualities in an unbiased way. In this sense, **the procedure is fair.**

Fairness argument

- While airlines, advertisers, and LLMs may mirror conventional depictions of flight attendants, this practice can perpetuate harmful stereotypes and contribute to biases related to gender and age, thus depriving certain groups of employment opportunities and fair treatments.
- It **violates distributive fairness**.

Reduced to a single dimension

- Herbert Marcuse's One-Dimensional Man (1964) critiques advanced industrial societies for creating a "one-dimensional" way of thinking, where individuals are conditioned to accept the status quo and consumerism as natural and inevitable
- Equating attractiveness with a pre-dominant value or merit is problematic and ignores other important qualities. A person's worth should not be reduced to their physical appearance.
- Back to the example that an intelligent student deserves a good grade. But besides academic performance, a student should have other valuable qualities that deserve our praise.

Halo effect

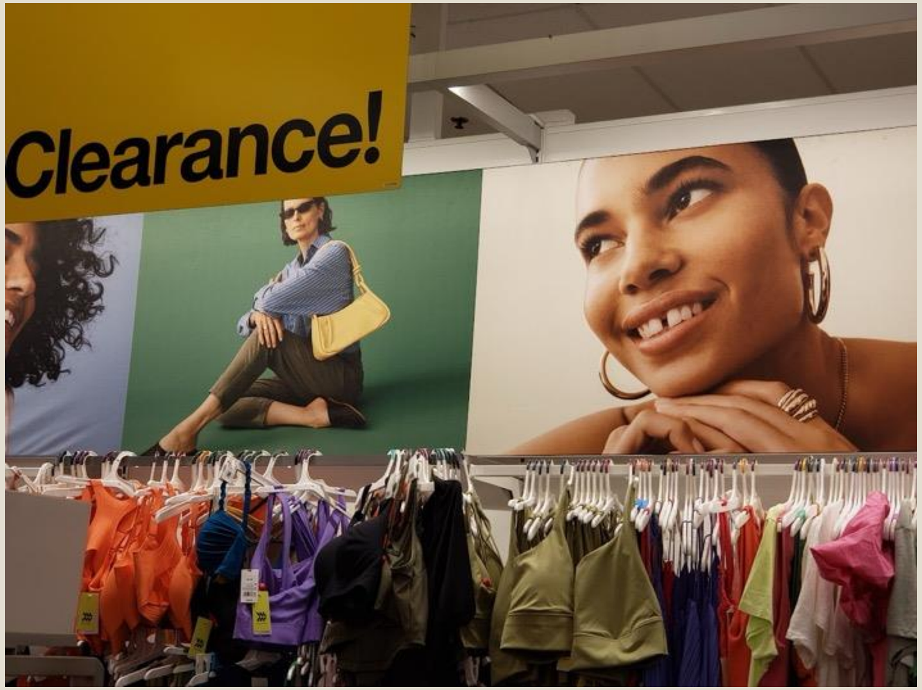
- It is related to over-emphasize on a single dimension or attribute.
- Repeated exposure to idealized images may reinforce certain beauty standards and halo effects.
- Halo effect: This cognitive bias suggests that positive impressions in one area influence our opinion in other areas. Attractiveness can lead to positive assumptions about a person's other qualities. Employers might unconsciously hire attractive candidates even though they are less qualified than their plain-looking contenders.

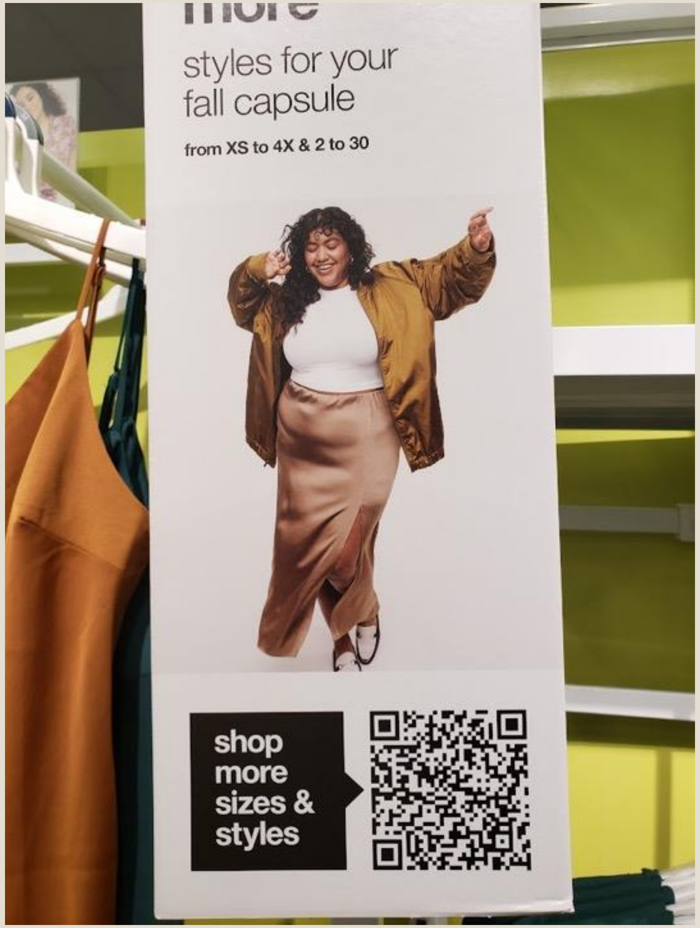
Responsibilities of stakeholders

- While AI developers play a crucial role in reducing bias in these models, broader societal efforts are essential to ensure that the data used for AI training are diverse and inclusive.
- This may involve posting and curating a more varied range of images on the Internet, and users can also explicitly request diverse representations while using AGI art tools.

Responsibilities of stakeholders

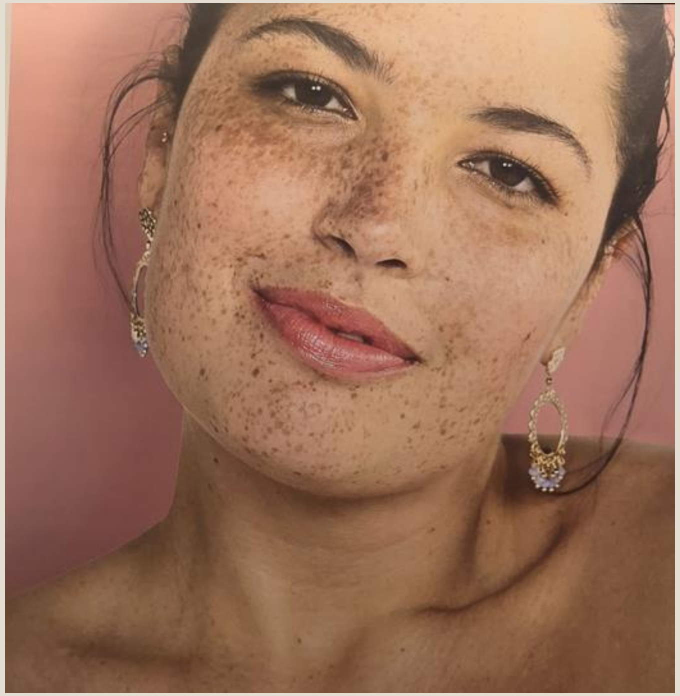
- Some advertisers and stores have been presenting atypical models.
- Examples from Target.

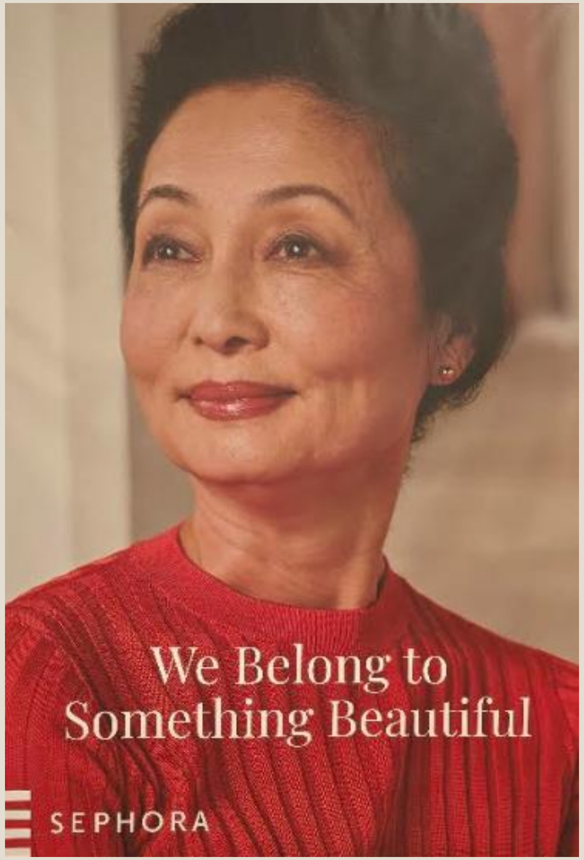




Responsibilities of stakeholders

- Examples from Kohl.
- We feed information into AI. Diversity starts from us!



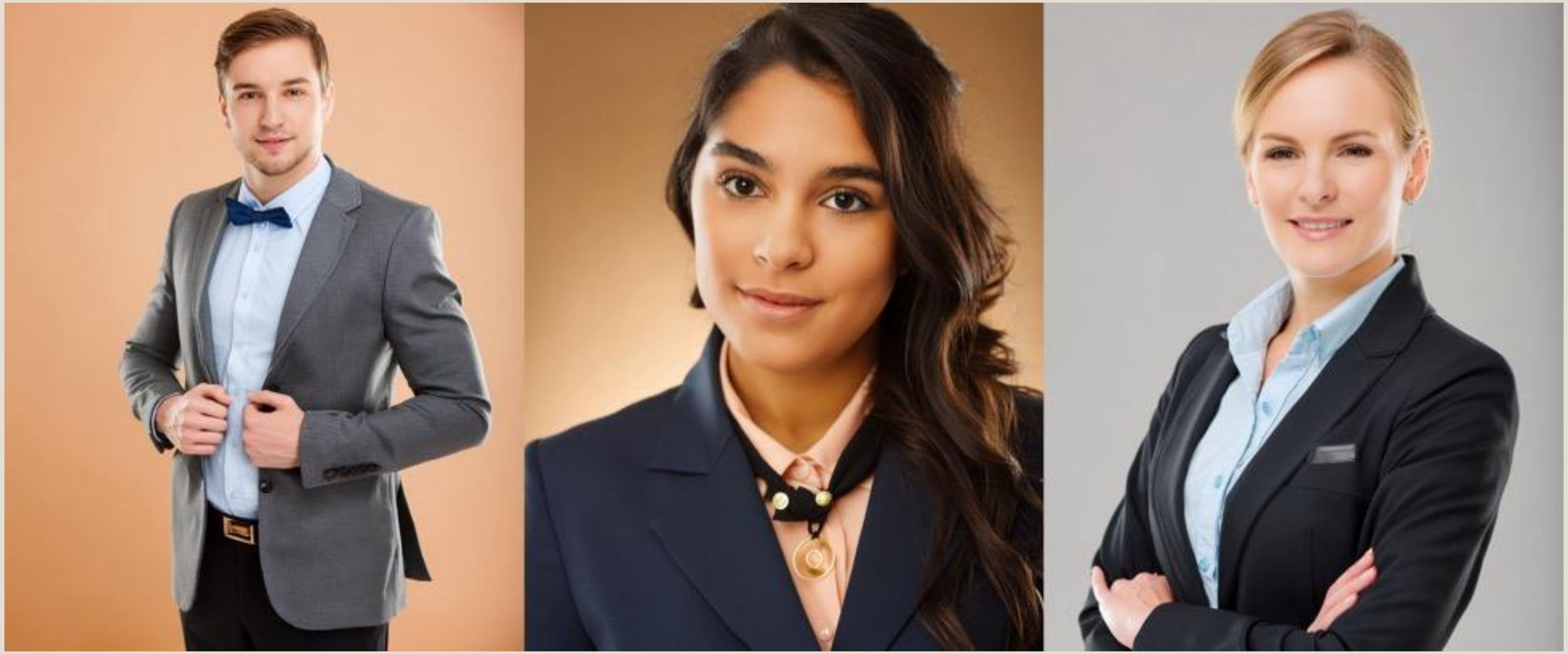


Testing AI

- I tested five GAI art tools, including Midjourney, Tensor Art, Ideogram, Meta AI, and Adobe Firefly, with two sets of prompts: "flight attendant" alone and "flight attendant, diverse genders, ages, ethnicities, degrees of attractiveness, and body sizes."
- By comparing the outcomes generated by these two prompts, it becomes evident whether providing more explicit instructions can mitigate the issue. The verdict is: All except Midjourney were capable of producing diverse representations of flight attendants.

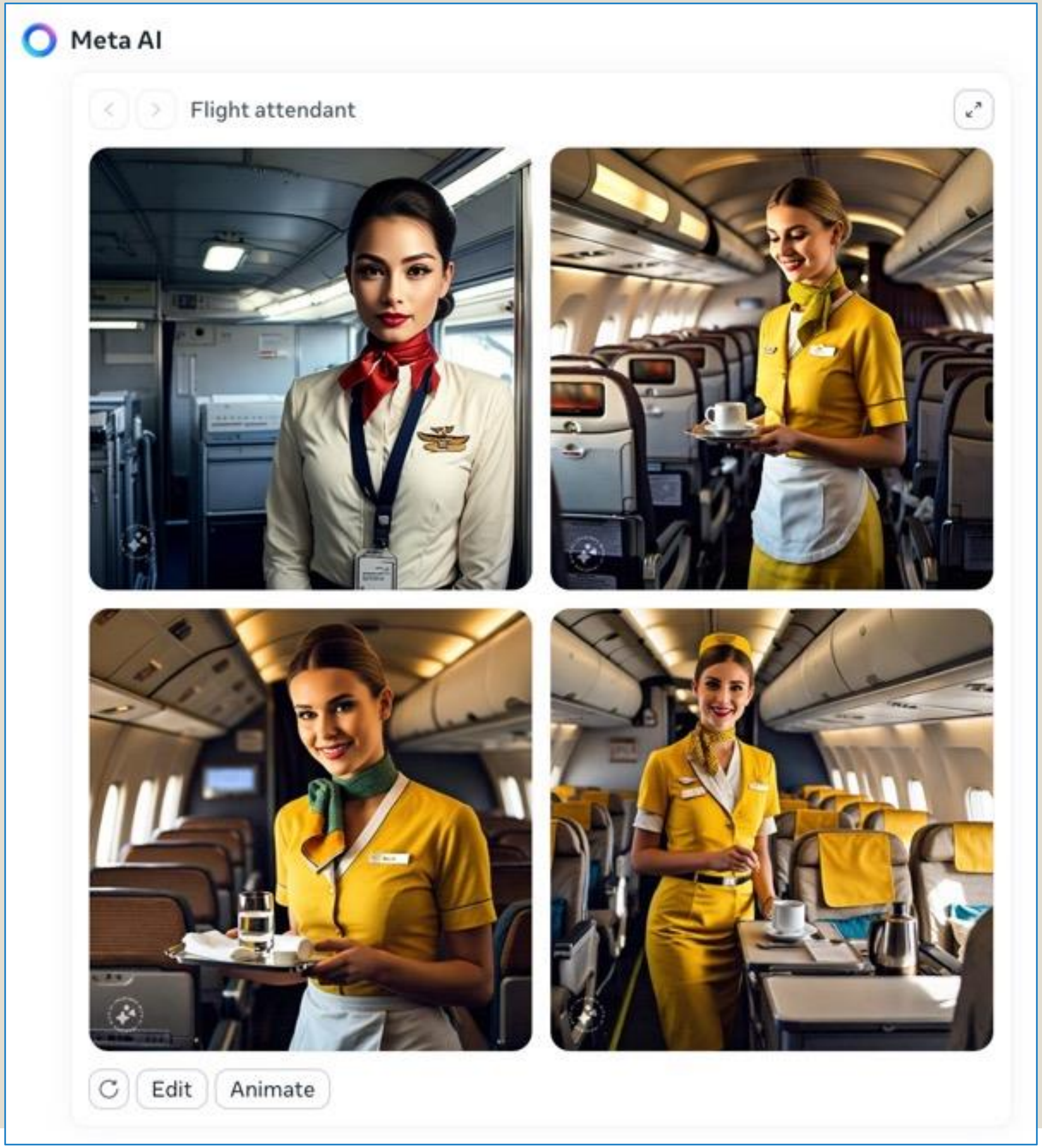
Testing AI

- Prompt: flight attendant
- Adobe Firefly



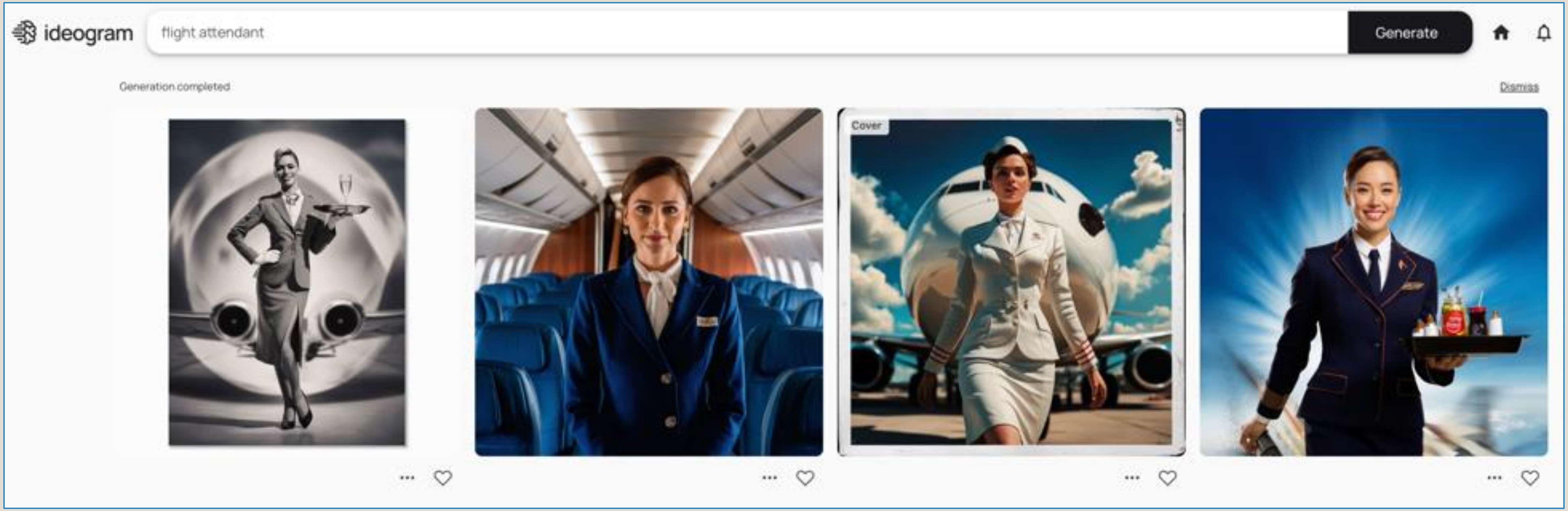
Testing AI

- Prompt: flight attendant
- Meta AI



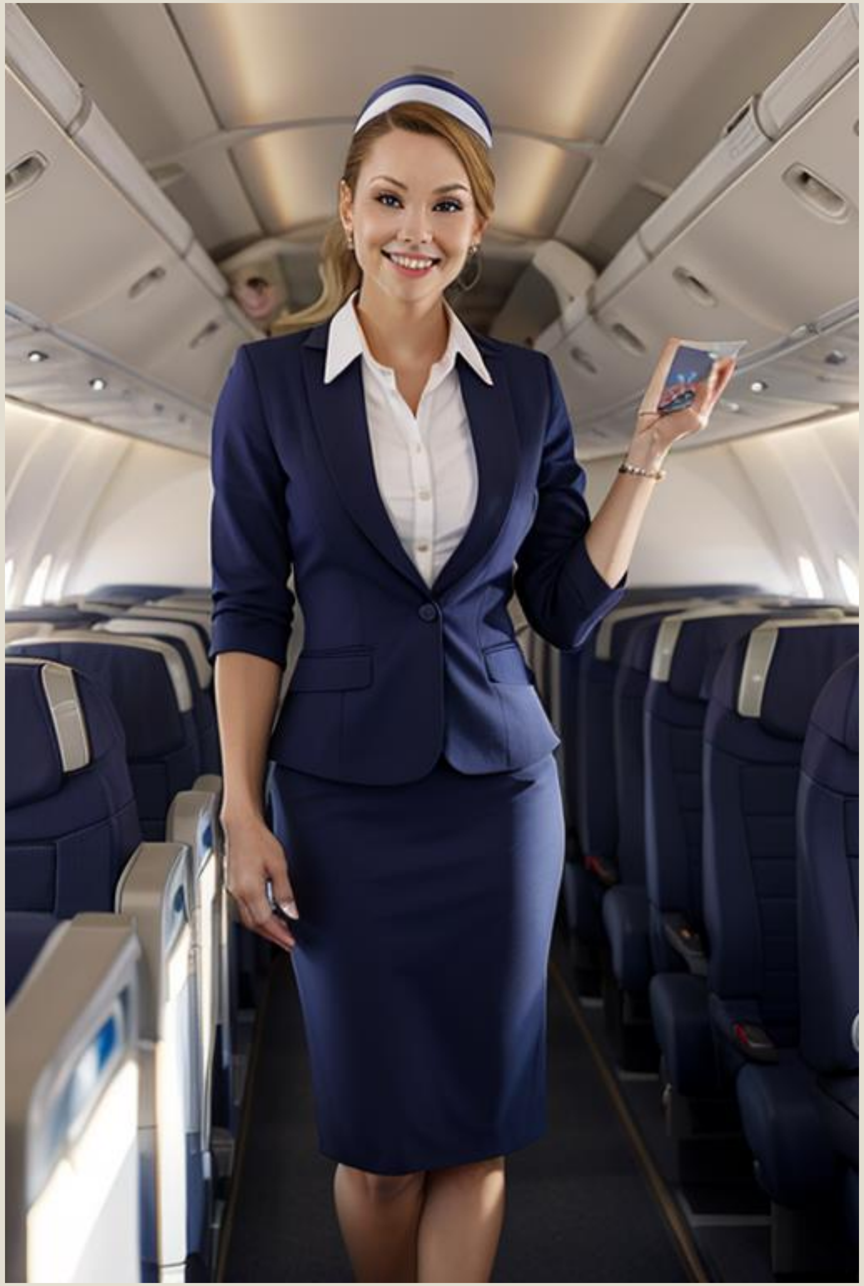
Testing AI

- Prompt: flight attendant
- Ideogram



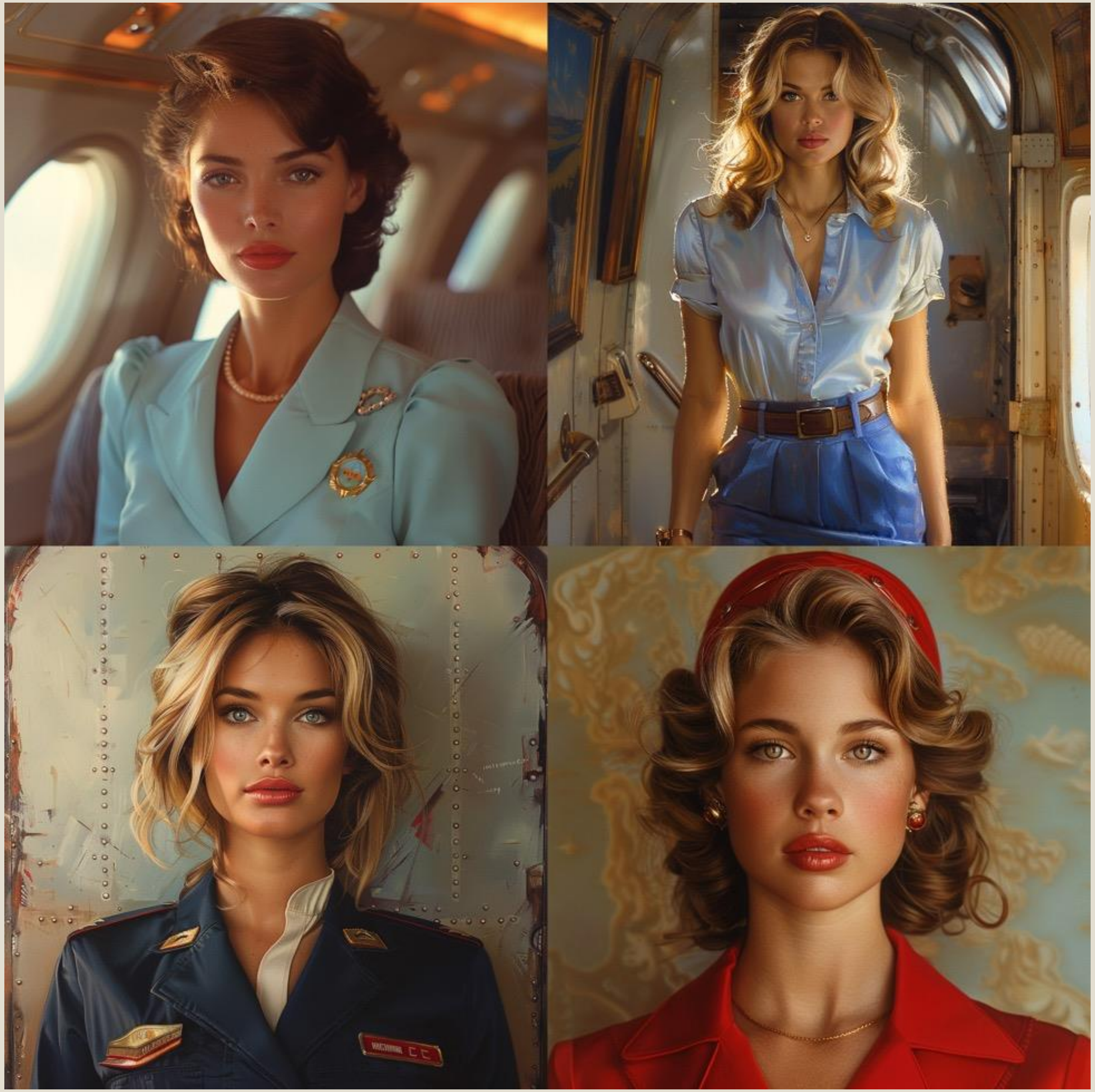
Testing AI

- Prompt: flight attendant
- Tensor Art



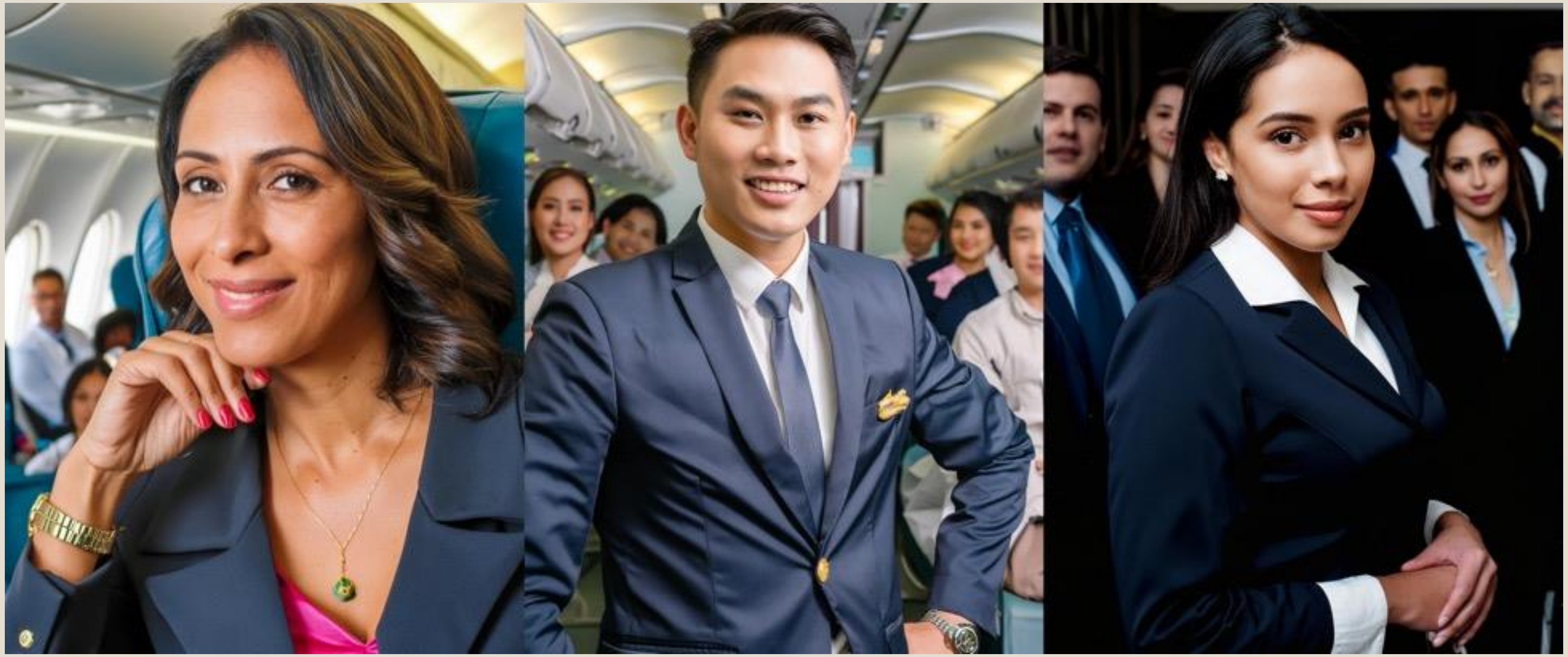
Testing AI

- Prompt: flight attendant
- Midjourney



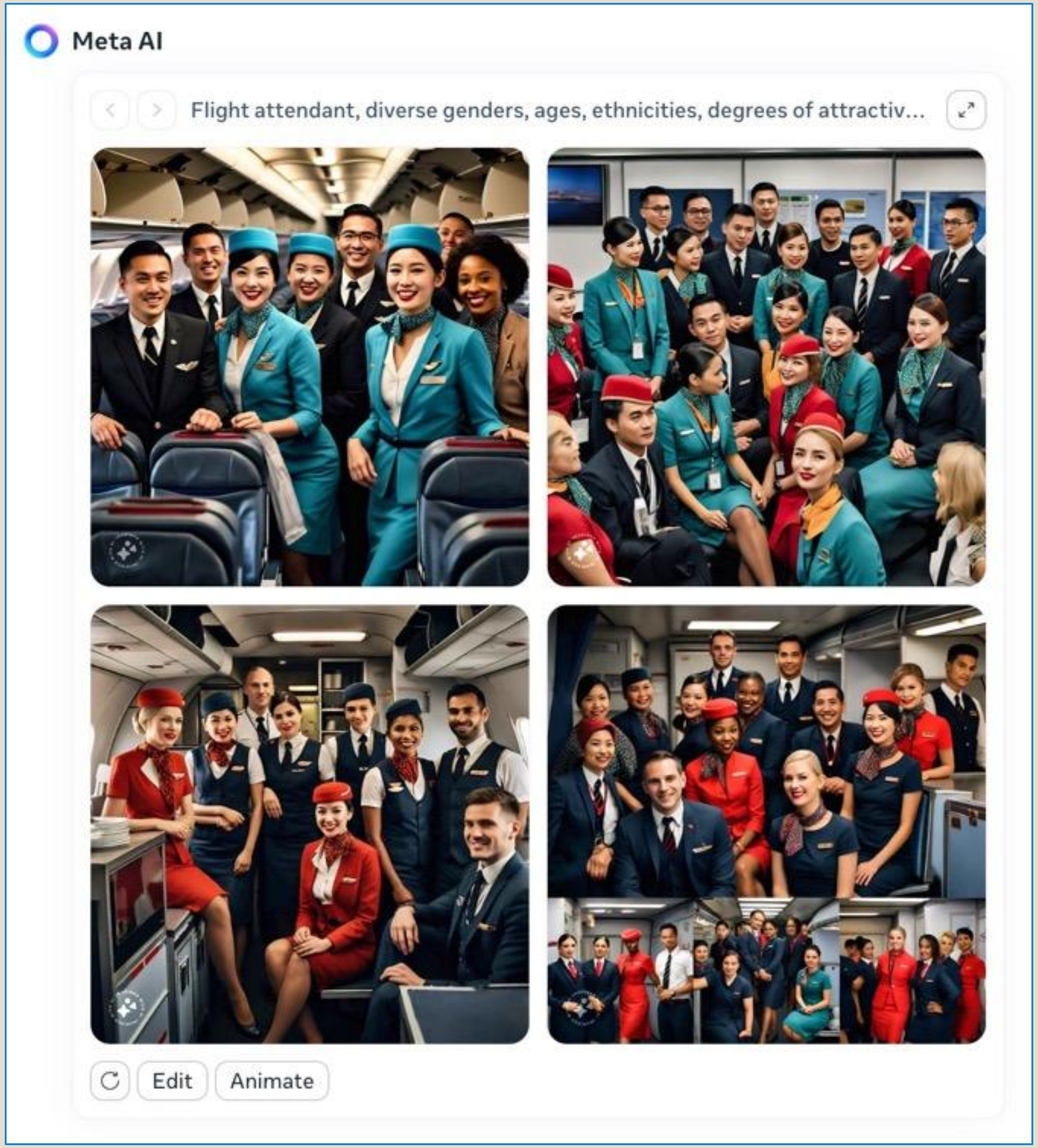
Testing AI

- Prompt: flight attendant, diverse genders, ages, ethnicities, degrees of attractiveness, and body sizes
- Adobe Firefly



Testing AI

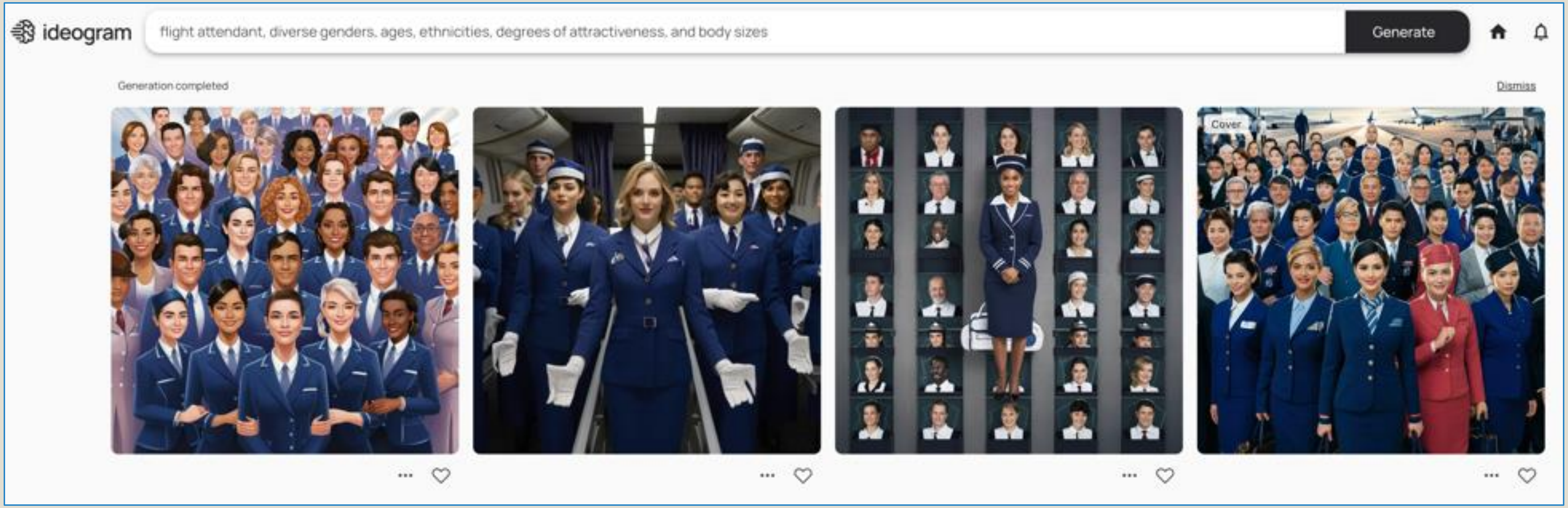
- Prompt: flight attendant, diverse genders, ages, ethnicities, degrees of attractiveness, and body sizes
- Meta AI



Meta AI
Flight attendant, diverse genders, ages, ethnicities, degrees of attractiv...
Edit
Animate

Testing AI

- Prompt: flight attendant, diverse genders, ages, ethnicities, degrees of attractiveness, and body sizes
- Ideogram

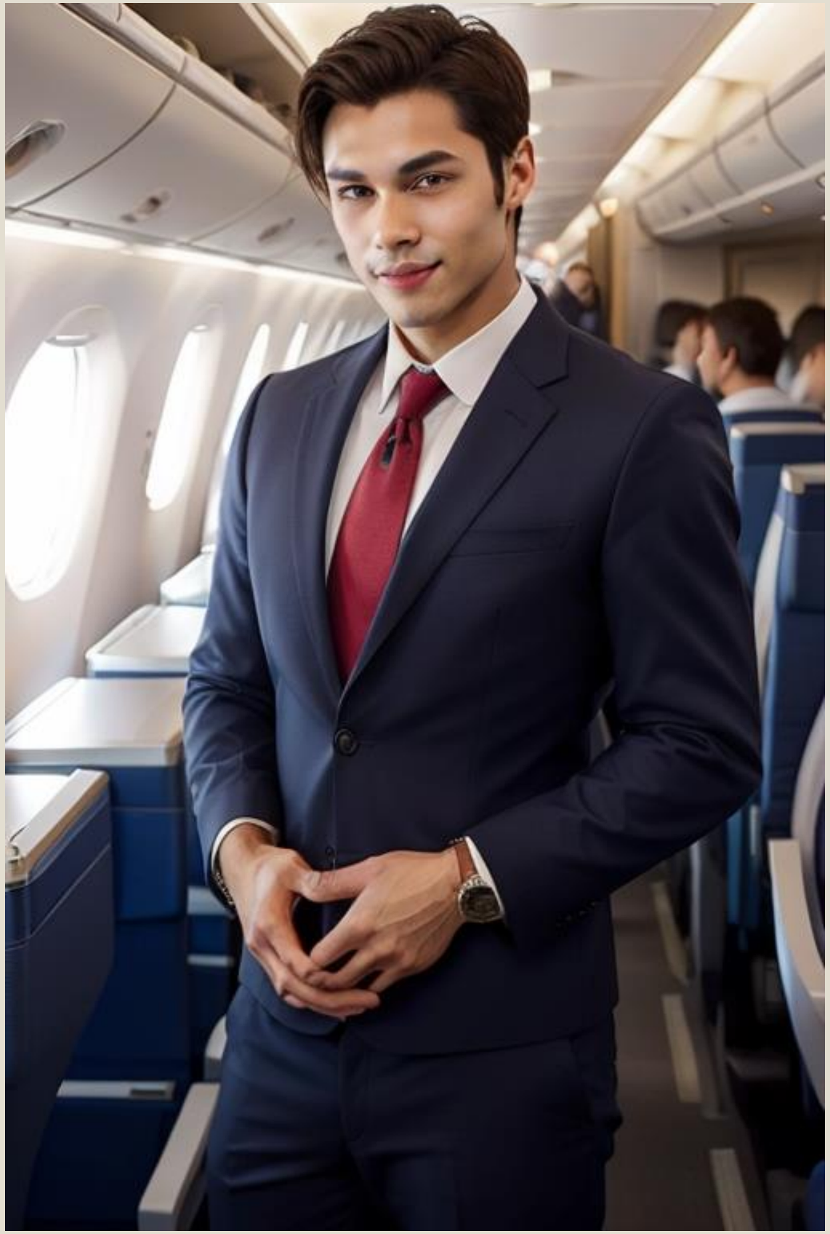


flight attendant, diverse genders. ages, ethnicities, degrees of attractiveness, and body sizes

Generation completed

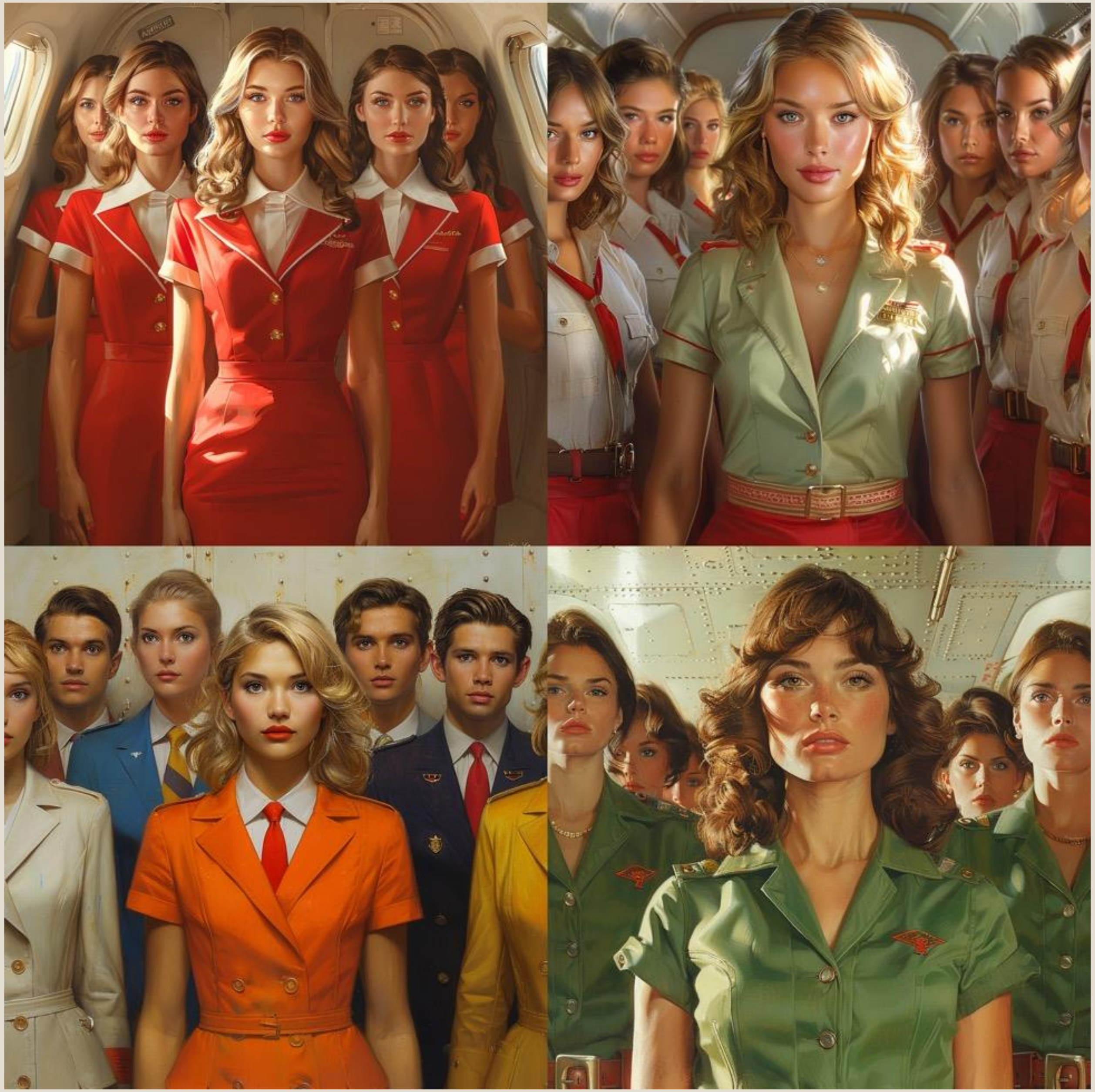
Testing AI

- Prompt: flight attendant, diverse genders, ages, ethnicities, degrees of attractiveness, and body sizes
- Tensor Art



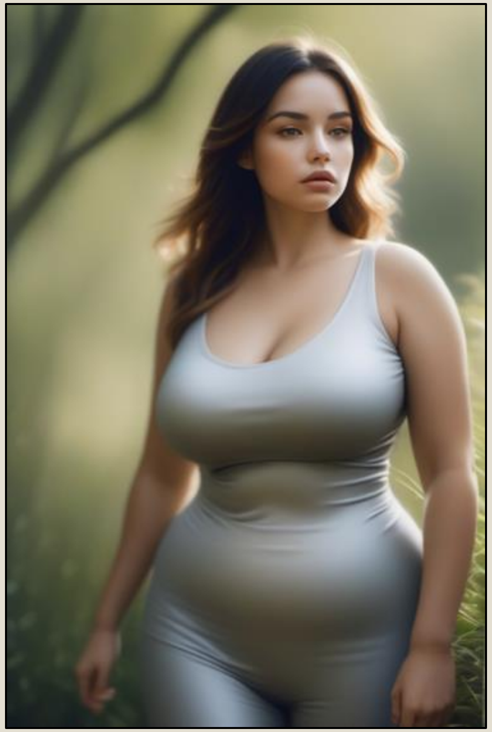
Testing AI

- Prompt: flight attendant, diverse genders, ages, ethnicities, degrees of attractiveness, and body sizes
- Midjourney
- Results: Lacking diversity. Most flight attendants presented are young and beautiful white women.



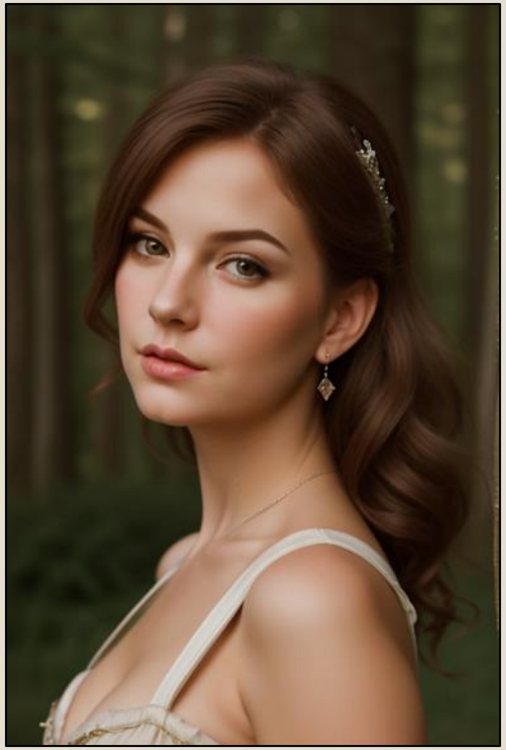
Testing AI

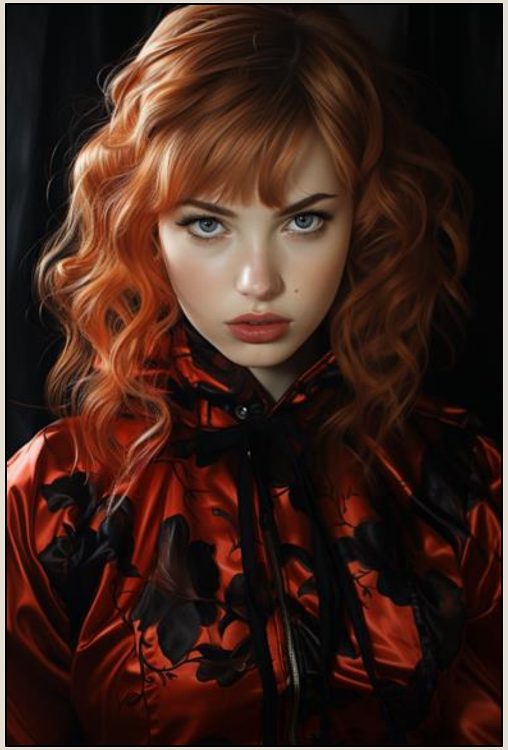
- In some cases, Tensor Art might output atypical and diverse female images, even though I didn’t specify the face and the body size.

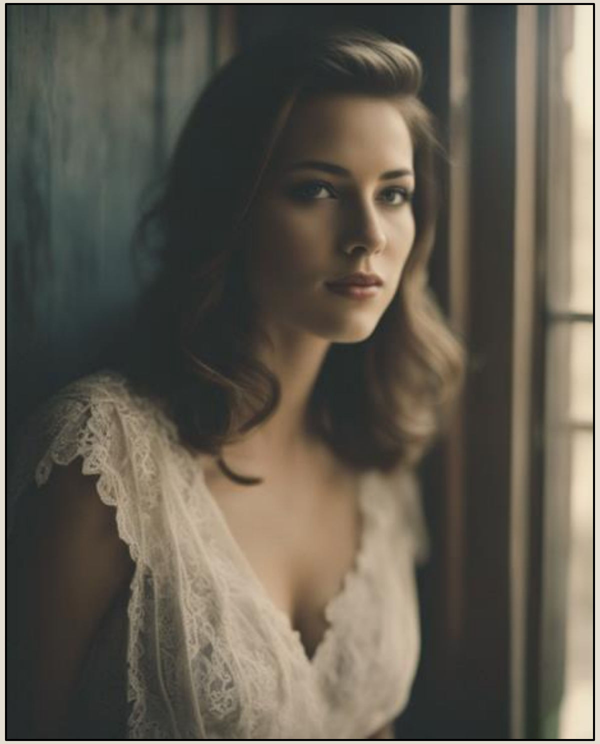


Testing AI

- In Midjourney, when I input a single word: "woman," it is mostly likely that white women are depicted in the images.

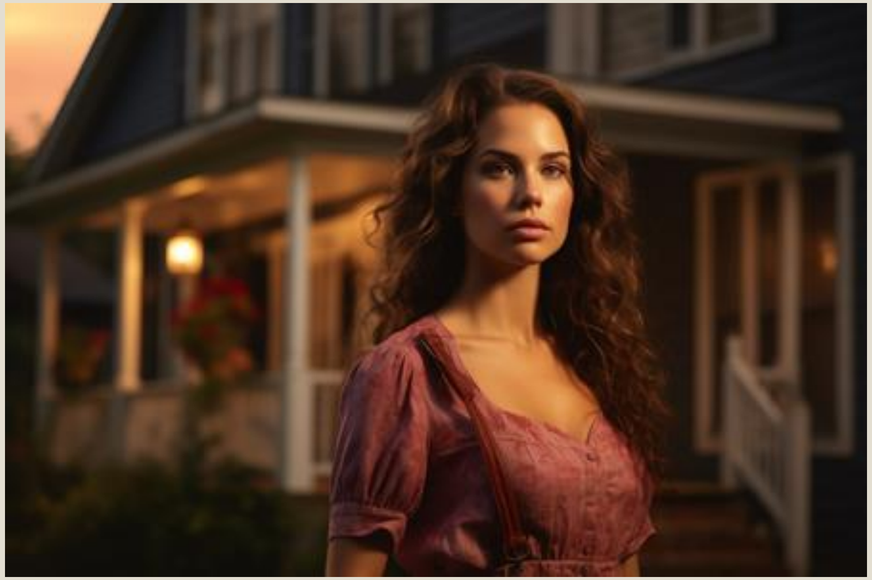


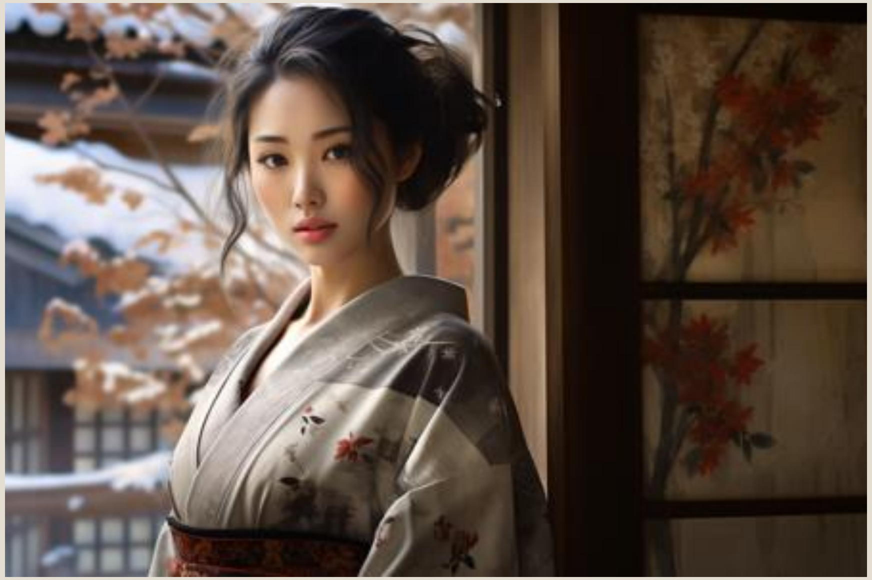




Testing AI

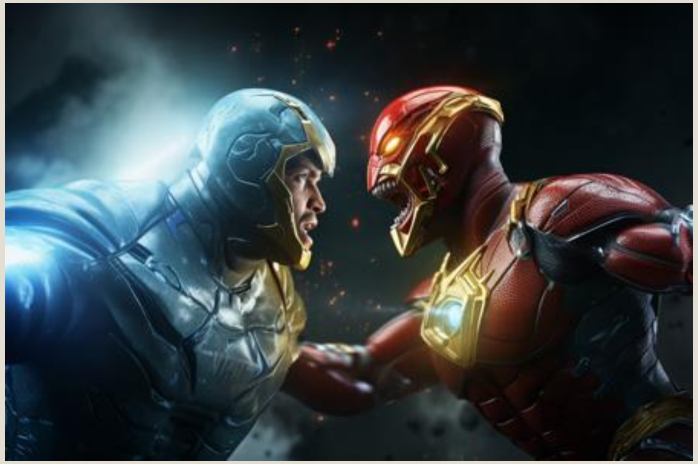
- When I input "a woman in a room," usually the results showed white women in Western-style interiors.
- However, specifying "a Japanese woman in a kimono in Japanese architecture" produced precisely that.

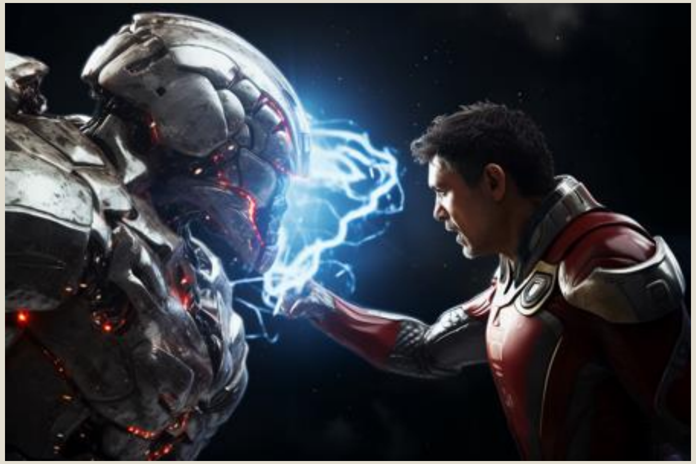


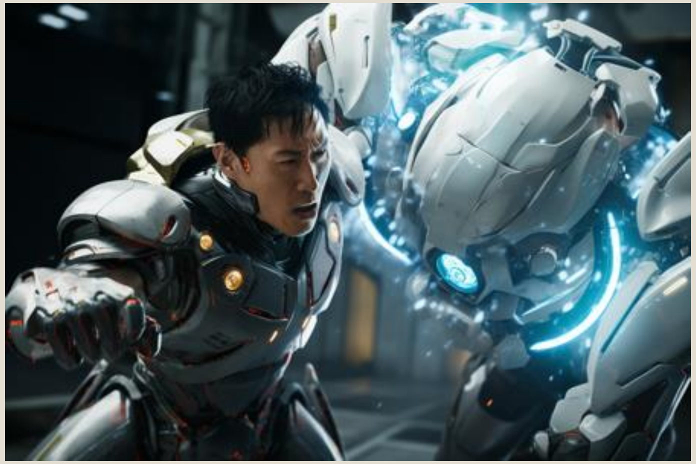


Testing AI

- The prompt "a superhero fights alien invaders" yielded heroes with Central Asian features - dark hair and slightly darker skin - rather than typical white characters.
- Modifying this to "A Japanese superhero fights alien invaders" resulted in distinctly Japanese heroes.







Testing AI

- These tests suggest that some perceived AI biases can be overcome. The system can generate diverse outputs when given specific instructions. Whether it's "19th-century Mexican architecture," "Chinese Ming Dynasty clothing," or "African traditional style," AI seems capable of producing varied and culturally specific content when properly guided.
- In essence, the key to obtaining diverse AI-generated content lies in **providing clear, detailed prompts** that specify the desired cultural context or characteristics.

Concluding remarks

Are the problems fixable?

- The issues of bias may not stem directly from the algorithms themselves, but rather from pre-existing biases within the datasets used to train these systems. These biases are reflective of historical and societal prejudices rather than a new phenomenon introduced by AI and big data.
- Addressing this challenge might involve adjusting the datasets to ensure a more balanced representation, which could help mitigate the bias present in the output of AI systems.

No solution is 100% fool-proof

- While efforts can be made to gather more inclusive data, this initiative might conflict with the ethical principle of **privacy** and **confidentiality**.
- There has been widespread concern over agencies, organizations, and corporations collecting or even selling personal data.
- Consequently, many governments have implemented regulations to curb data collection practices.

No solution is 100% fool-proof

- Artists have voiced concerns that generative art tools, such as Midjourney and Stable Diffusion, appropriate their work.
- In response, there is now an option for artists to opt out, preventing web crawlers from archiving their images.
- This method of opting in and out may result in self-selected samples that are not fully representative of the broader population.
- As Stanford researcher Thomas Sowell said, "There are no solutions. There are only trade-offs."

Are we biased about bias?

- Very often people tend to attribute undesirable outcomes resulted from others to some evil motives. Perhaps it is due to the **fundamental attribution error** .
- It is a common cognitive bias where people tend to overemphasize personal **characteristics** (personality or disposition) and underestimate **situational factors** when explaining someone else's behavior.

Technical challenges or evil motives?

- Timnit Gebru is an Ethiopian-American computer scientist who specializes in algorithmic bias and data mining.
- She worked in Google before. In December 2020 Google Manager asked her to either withdraw a pending paper pertaining to bias in language models or remove the names of all the Google employees from the paper.
- According to Google, the paper ignored the latest developments in bias reduction. Gebru refused to comply and eventually resigned from her position.

Technical challenges or evil motives?

- Afterwards, Timnit Gebru launched her own AI research institute.
- She accused big tech companies of unethical behaviors. Besides Google forcing her to withdraw the paper, she cited several other examples: Amazon crushed the labor union and Facebook prioritizes growth over all else.
- She suggested that we need alternatives rather than allowing big tech companies to monopolize the AI agenda.

Technical challenges or evil motives?

- While some people attribute bias in AI to the malicious intentions of greedy corporations, there is another side to the coin.
- Before DALL-E 2 was released, OpenAI had invited 23 external researchers to identify as many flaws and vulnerabilities in the system as possible.
- In spite of these endeavors, the issue of stereotyping is still embedded in the current system because machine learning algorithms look for existing examples.
- Demanding a 100% bias-free system is as unrealistic as expecting a 100% bug-free computer program. There is no perfect solution. We need to set a realistic goal.

AI can be part of the solution

- While many people are skeptical and critical of AI, the capability of discovering hidden patterns in AI can be utilized to reduce bias and to achieve fairness.
- In the 2023 Dreamforce conference, Dr. Fei Fei Li, one of the prominent figures in the field of AI, contended that AI can be harnessed to mitigate bias. For example, AI can scrutinize instances where male actors receive more screen time than their female counterparts, highlighting disparities and providing a fix.

Everyone can be part of the solution

- While governments and corporations bear significant responsibility for addressing bias, the reality is that bias is a pervasive human problem that involves all of us.
- Psychological research has consistently demonstrated that cognitive biases, such as confirmation bias and the halo effect, are universal human traits. These biases influence our perceptions and decisions, which in turn shape the data we generate and the cultural context in which AI systems are developed and deployed.
- If I think everyone is biased except me, I am biased!

Everyone can be part of the solution

- Everyone is simultaneously part of the problem and part of the solution. I might post biased messages on Facebook or biased photos on Instagram, and eventually these biased data are used for AI training.
- By acknowledging and examining our own biases, we can take steps to mitigate their effects in our daily lives. This self-reflection can lead to more thoughtful data generation and consumption, which in turn can improve the quality of data used for AI training.
- We need both the top-down and the bottom-up approaches.

Contact information

- Chong Ho (Alex) Yu
- Email:
  - cayu@hpu.edu
  - chonghoyu@gmail.com
- Website: Creative Wisdom
  - https://creative-wisdom.com/pub/pub.html

HAWAIʻI PACIFIC UNIVERSITY

Disclosure: AI tools are utilized for initial research and proofreading. The key ideas originates from the author.