The Algorithmic Shade Room

An experiment on how bias in curated social media feeds shapes trust and platform fairness

Student Name:	Zahra Kraag
Student Number:	558813
Supervisor:	Prof. Dr. Marc Verboord

Master Media Studies - Media & Creative Industries
Erasmus School of History, Culture and Communication
Erasmus University Rotterdam

Master's Thesis
June 2025

Word Count: 12677

The Algorithmic Shade Room

An experiment on how bias in curated social media feeds shapes trust and platform fairness

ABSTRACT

As algorithms increasingly shape the way individuals consume digital content, many scholars have taken it upon themselves to investigate how these systems may reinforce social inequalities. One specific and often overlooked manifestation of algorithmic bias is colorism, the preferential treatment of individuals with lighter skin tones. While previous research has examined algorithmic discrimination in areas such as healthcare, policing, and education, less is known about colorist bias embedded in social media feeds and how it can affect users' perceptions. This thesis investigates the extent to which colorist algorithmic biases influence perceptions of platform trust and fairness, and whether comfort with racial diversity in friend groups moderates this relationship. A between-subject survey experiment was conducted with 216 participants. Each participant was randomly assigned to view a mock Instagram feed curated with either predominantly lighter-skinned individuals or a racially diverse set of individuals. Participants completed a pre- and post-test measuring their perceived trust and fairness toward the platform, as well as a scale measuring the racial diversity of their social circles. The findings revealed that exposure to racially diverse Instagram feeds significantly increased perceptions of platforms trust and fairness. In contrast, exposure to feeds with lighter-skinned individuals did not significantly decrease perceptions of trust and fairness. The moderation analysis revealed comfort with racial diversity in friend groups did not significantly influence the relationship between algorithmic bias and user perceptions. These results demonstrate that visible racial diversity in content matters for how platforms are perceived, regardless of users' cross-group friendships. The findings of this thesis contribute to ongoing discussions in algorithmic fairness by highlighting the psychological impact of visual diversity in algorithmically curated content. The findings also introduce colorism as a critical variable in the broader discourse of digital equity and representation, specifically in personalized recommendation systems. Even though the study was limited by factors such as ecological validity and the sensitivity of topics such as race and discrimination, the implications highlight the importance of inclusive algorithmic design and enhanced transparency in AI systems. By holding platforms accountable for their unequal values embedded in their algorithms, this research advocates for more ethical, inclusive, and socially aware technology designs.

KEYWORDS: *Algorithmic bias, fairness, trust, colorism, social media*

Table of Contents

1. Introduction

It was September 2020, during the height of the COVID-19 pandemic. Robin Aisha Pocornie was getting ready behind her desk to take an online proctored exam. When she reaches the facial recognition step of the log in process, the software failed to detect her face and displayed the message: “Face not found. Room too dark”. She tried again, but it failed once more. As she was running out of time to log in, she was forced to shine a lamp directly into her face; only then did the system recognize her face and allow her to access the exam. None of her lighter-skinned classmates encountered this issue when logging in. Confused and unsettled, Pocornie began researching similar experiences and noticed a pattern; it turns out that many people of color have also struggled with algorithmic systems failing to recognize their face. With the help of the Amsterdam Race & Technology Center, she filed a complaint with the university board of directors and the exam commission and calls out the discriminatory effects of facial recognition software used in education (TEDxTalks, 2024).

Pocornie's experience is just one of several that point to a more serious, fundamental flaw in the way artificial intelligence systems function. This is not only present in facial recognition software programs, but skin-type biases are present in all types of commercial artificial intelligence systems. For instance, commercial algorithms in healthcare systems can unfairly assign fewer medical resources to Black patients (Agarwal et al., 2022, p. 2). This exemplifies algorithmic bias, a socio-technical phenomenon that negatively affects marginalized or underprivileged groups, as these are manifestations of social biases that favor certain groups or individuals (Kordzadeh & Ghasemaghaei, 2022, p. 390). Technology can have violent impacts on marginalized individuals, but the discriminatory nature of it only worsens this (Hampton, 2021, p. 2). An example of this is predictive crime-mapping tools which are used by police to patrol high crime neighborhoods (Kordzadeh & Ghasemaghaei, 2022, p. 390). These systems utilize historical data, which includes crime incidents reported by police officers that disproportionally target areas with underprivileged populations and racial minorities, thus worsening existing inequalities.

It is evident that these biases are rooted in colorist ideologies. Colorism is the differential treatment of people based on skin color (Trammel, 2023, p. 51). It is a form of discrimination in which lighter-skinned people are favored more over dark-skinned people. Slavery and European colonialism established the foundation of colorism in large parts of the world (Hunter, 2012, p. 247). Colorism is largely prevalent in areas such as income, education, criminal justice sentencing, housing, and the marriage market. However, in recent times social media algorithms are starting to favor lighter-skinned people over dark-skinned people (Ryan-Mosley, 2021, par. 12). For example, Love (2022, p. 115) found that content that contains lighter-skinned individuals receive more engagement than dark-skinned individuals. The silencing of marginalized groups is called *shadowbanning*, where platforms prevent content from certain individuals appearing in the searches without their knowledge (Jarvis & Quinlan, 2022, p. 137).

Another example is the filters on social media platforms such as Instagram that contribute to colorist practices by promoting Eurocentric beauty standards (Trammel, 2023, p. 60). Due to these filters, colorism continues to perpetuate in the beauty industry through skin bleaching products. Countries in Africa, Europe, and Asia have banned these filters and skin bleaching products because they are deemed to be too dangerous (Trammel, 2023, p. 61). According to Baha (2023, p. 3), biased algorithms result from tech companies led by White men which are set up to reflect their ideology, often neglecting marginalized groups. This is exemplified in the harmful algorithmic outcomes when searching in Google search for “black girls”; the algorithm would return explicit pornographic content (Noble, 2018, p. 66). There is a lack of diversity and cultural awareness in tech industries about stereotypical portrayals of black women (Noble, 2018, p. 70). This lack of diversity is one of the main sources of algorithmic bias and can arise in all types of digital environments. Consequently, digital platforms and their recommendation algorithms reinforce existing biases by favoring whiteness and limiting the visibility of marginalized groups.

Existing societal biases might manifest themselves in the algorithm’s output, which is another major source of algorithmic bias. Inaccurate or poorly selected input data can also result in algorithmic bias (Kordzadeh & Ghasemaghaei, 2022, p. 390). Ideally, algorithms are supposed to be neutral; they follow fixed rules and use data to make decisions in an objective way to aim for fair and accurate results without bias (Jacoba et al., 2023, p. 433). But in reality, under-represented trained datasets may not accurately reflect marginalized groups. Some potential causes of these algorithmic biases include prejudice, structural bias, choice of training data, and many more (Woodruff et al., 2018, p. 2). For instance, Jacoba et al. (2023, p. 434) found that 45% of the global population does not have readily accessible representative ophthalmic images (medical images related to eye health and vision care). This can result in the misdiagnoses or misrepresentation of several retinal diseases in marginalized groups. The AI models are trained with data from one ethnic group and may not accurately diagnose members from other ethnic groups, thus increasing health inequalities (Jacoba et al., 2023, p. 435).

Such examples highlight how algorithmic bias can influence real-world outcomes, particularly for underrepresented communities. Aside from physical impacts, algorithmic bias can also influence how people perceive the fairness of algorithmic systems in their daily lives. Chen & Sundar (2024, p. 4) argue that a lack of diversity in training data can greatly affect perceived algorithmic fairness. In their study, participants were shown an AI hiring system that analyzed facial expressions with varying conditions of racial diversity in training data and labelers. Their finding revealed that transparency and racial diversity in an algorithm can enhance user trust and perceived fairness. Hence, trust in a platform increases when users view its algorithm as fair (Wu et al., 2023, p. 7). While the algorithm may function adequately, it could be deemed unfair if certain groups benefit more. Social media platforms curate content using algorithms, and biased algorithms may make the platform seem unfair.

With the abundance of literature focusing on how algorithmic bias can manifest in areas such as health care, hiring, policing, facial recognition, and search engines (Agarwal et al., 2022, p. 2; Kordzadeh & Ghasemaghaei, 2022, p. 390; TEDx Talks, 2024; Noble, 2018, p. 69), and the recent attention placed on race and fairness in technology and algorithms (Noble, 2018, p. 69; Love, 2022, p. 115; Woodruff et al., 2018, p. 2; Lu et al., 2024, p. 812; Trammel, 2023, p. 60), this thesis aims to contribute to this field by focusing on how the role of algorithmic bias and colorism plays in curated social media content. As social media plays a big role in our everyday life and influences our attitudes and beliefs, this research looks at how colorist algorithmic biases on social media curated content can influence user perceptions of trust and fairness in platforms. This research poses the following question: *To what extent do colorist algorithmic biases impact users' perceived trust of and platform fairness of curated social media feeds?*

To answer this research question, a between-subject experimental design will be used to test whether biased algorithmic Instagram feeds will affect perceptions of trust and fairness of the platform. The experiment will have two conditions, consisting of curated Instagram feeds with predominantly lighter-skinned individuals and with individuals of diverse skin tones. Participants will be randomly assigned to one of these conditions. Furthermore, this thesis will also test whether having a diverse friend group strengthens the relationship between biased algorithmic feeds and trust and fairness. This thesis will start off with a literature review, which will pose as theoretical grounding for this research and examine existing literature to form hypotheses. Followed by that will be the methods section, in which the methodology choice and measurements will be justified, and the results section. The thesis will end with a discussion, in which the main findings of the study will be evaluated, limitations, and suggestions for future research.

1.1 Societal relevance

The societal relevance of this thesis addresses the broader impact of biased algorithms and contributes to advocating for ethical algorithmic design. As fairness is often closely related to discrimination, evaluating and scrutinizing fairness of algorithms plays an important role in shaping the lives of users, especially those in minority groups (Chuan et al., 2024, p. 11). The reinforcement of colorism and systematic inequalities is detrimental for marginalized communities, as it limits their visibility and representation (Hampton, 2021, p. 4). This is particularly concerning on social media platforms, as visibility and recommendation algorithms play a big role in shaping public perception, social inclusion, and economic opportunities for users and creators.

Moreover, as algorithms increasingly mediate the kind of content we consume, how much, and how often, there is a growing possibility that marginalized groups will remain excluded or misrepresented. This not only reinforces offline societal hierarchies but also introduces new forms of digital inequalities. Algorithmically curated content that favors lighter-skinned individuals perpetuates standards and beliefs tied to colonial histories, which can have real-world consequences on their self-esteem, mental health, and access to opportunities for people of color (Hern, 2021, par. 7; Ryan-Mosley, 2021, par. 14; Keyes et al. 2020, p. 692; Hampton, 2021, p. 4). This thesis seeks to promote diversity, inclusiveness, digital justice, and awareness while holding platforms accountable. It ultimately seeks to add to the existing discourse around digital justice and advocates for transparency, diversity, and ethical practices in the development and implementation of algorithmic systems.

1.2 Academic relevance

Previous literature surrounding algorithmic discrimination is largely explored in healthcare, digital advertising, policing, or search results (Baha, 2023, p. 6; Lu et al., 2024, p. 812; Noble, 2018, p. 69; Agarwal et al., 2022, p. 2). Many of these works offer a solid theoretical foundation for this thesis, however, there remains a notable gap in research addressing how colorism specifically—as opposed to broader racial categories—plays a role in shaping user trust and fairness perceptions of social media platforms. Additionally, few studies have explored the psychological effects of exposure to biased visual content. This thesis will add a unique contribution to the subject of algorithmic fairness—particularly in the realm of recommendation systems and platforms like

Instagram—and it extends scholarship on algorithmic bias by also testing the role of cross-group friendships and adds depth to ongoing studies about algorithmic transparency and bias mitigation. Additionally, this thesis offers an interdisciplinary perspective to understanding algorithmic colorism. It offers empirical and theoretical insights into algorithms in digital spaces by operationalizing trust and fairness as outcomes of exposure. Lastly, the experimental design of this research offers methodological contribution and paves way for new methods to study biased algorithms.

2. Theoretical Framework

2.1 Algorithmic bias and fairness

In computer science and mathematics, an algorithm is a collection of procedures used to solve or analyze a problem (Baha, 2023, p. 1). They are mostly used for data processing, calculating, and sorting. Typically, they receive an input, use them to perform a series of steps, and then output the results. (Baha, 2023, p. 1). In the digital sphere, algorithms are commonly used to rank, filter, and recommend content using the users' information. This is exemplified by the recommended content that users receive in their social media feeds. Such recommendations are based on the types of content the user interacts with, the video information of the post, and the user's account settings such as language and location. Based on this, policies and reward systems help algorithms learn which content to prioritize based on the users' online activities with the goal of maximizing long-term payoffs (Calvano et al., 2024, p. 1).

Algorithmic bias, however, is the bias derived from algorithms. This can be defined as an algorithm consistently producing results that differ unfairly from an expected standard (Fazelpour & Danks, 2021, p. 2). Fazelpour and Danks (2021, p. 3) claim that algorithms are anything but objective or free from biases and thus argue that algorithms implement values that can introduce biases in two ways. First, depending on the domain an algorithm is programmed to function in, algorithms can unintentionally and unjustly favor one group over another. Second, algorithms are optimized for specific performance metrics. Algorithm developers play a key role in determining the implemented values that algorithms will take on when selecting particular performance metrics (Fazelpour & Danks, 2021, p. 3). The latter half of the two are viewed as one of the major sources of algorithmic bias.

According to the findings of Akter et al. (2021, p. 4-5), algorithmic biases can result from an underlying dataset, insufficient methodological approaches, or embedded societal factors. In the context of training datasets, sample selection bias could alter algorithmic decisions; this entails that a dataset fails to accurately represent a random sample from a target group. Furthermore, developers are likely to identify members from incorrect sample units as their target population regarding certain characteristics such as attitudes or values (Akter et al., 2021, p. 5). An example of this is the abandonment of Amazon's AI recruitment tool as it discriminated against female applicants due to the lack of female data in the training dataset. Similarly, inadequate data collection could also introduce biases. For instance, using celebrity images for a facial recognition dataset might underperform as there are underrepresented groups in celebrity culture (e.g., people of color, women, etc.) (Fazelpour & Danks, 2021, p. 6).

Societal biases are the most frequently mentioned sources of bias in discourse surrounding algorithmic bias. Learning algorithms aims to produce models that reflect the statistics of historical data (Fazelpour & Danks, 2021, p. 6). Hence, algorithmic biases can occur as a result from real-world existing bias. The problematic nature of algorithmic bias could negatively impact marginalized communities and reinforce existing inequalities. For example, Yapo and Weiss (2018, p. 5366) pointed out that Flickr has reportedly been exhibiting racist results by associating images of animals or apes with dark-skinned people. Furthermore, algorithm bias could limit the visibility of users online. For instance, algorithms used in recommendation systems can prevent customers from discovering new products or services (Akter et al., 2021, p. 6). Hampton (2021, p. 4) stated that experiencing colorism or any form of discrimination from a machine is damaging for one’s psyche. She uses the example of a Black artist's work being less visible on Twitter because the platform’s preview feature favored lighter skin tones, cropping out Black subjects. This reduced engagement with their art, negatively impacting their visibility and potential income.

There continues to be a common belief that decisions made by automated algorithms are expected to be more fair or objective (Pessach & Shmueli, 2023, p. 1). However, this continues to be disproved based on the numerous sources of these biases that were highlighted previously. Regardless of these sources, it’s still quite challenging to fully understand where or how these biases are formed. The prediction model may be inherently biased due to the retention of prior biases in its training data, or it may be more deeply ingrained in the code (Pessach & Shmueli, 2023, p. 1). This is where algorithmic fairness comes into play. The concept of algorithmic fairness does not have a clear-cut definition, as its definition greatly depends on the context in which the bias is taking place. For instance, Chen & Sundar (2024, p. 4) define algorithmic fairness as everyone, regardless of race or other protected characteristics, is supposed to have an equal chance at favorable classification. Conversely, Woodruff et al. (2018, p. 2) refer to algorithmic (un)fairness as the unfair or discriminatory treatment of people through algorithmic systems or algorithmically assisted decision-making based on characteristics like race, gender, or sexual orientation.

According to Pessach & Shmueli (2023, p. 3), there are two types of discrimination. First is disparate treatment, which is intentionally treating someone differently because they belong to a protected class (a group legally safeguarded from discrimination based on characteristics such as race, gender, religion, age, disability, or sexual orientation); this is also classified as direct discrimination. The second is disparate impact, which entails causing disproportionate harm to individuals in a protected group, even if the policy appears to be neutral; this is also deemed as indirect discrimination. The growing topic of discussion surrounding this has resulted in scholars developing criteria to assess and measure the fairness of an algorithm. Wang et al. (2022, p. 3) mentions “fairness through unawareness”; a notion that assumes that an algorithm is fair unless it does not explicitly consider sensitive characteristics (such as race or gender) when making a decision. “Fairness through awareness” or “individual fairness” entails that individuals with identical characteristics, experiences, or preferences are equally classified by an algorithm (Wang et al., 2022, p. 5).

In contrast to these individual-level approaches, other fairness approaches focus on group-level outcomes. For example, fairness criteria such as disparate impact and demographic parity highlight that different demographic groups should receive the same positive outcomes by preventing systematic disadvantages. This suggests that algorithms trained on data that exclude sensitive attributes (such as those that clearly indicate whether someone belongs to a protected group) are less likely to cause direct, intentional discrimination — but they can still lead to unintentional bias, resulting in unequal outcomes across groups (Pessach & Shmueli, 2023, p. 3-4). Furthermore, other refinements of group fairness include measures such as equalized odds and equal opportunity. This evaluates whether different groups experience similar error rates or classification accuracy (Pessach & Shmueli, 2023, p. 4-5). These insights indicate that next to outcomes fairness should also be consistent in algorithmic performance. However, these approaches may be incompatible with individual fairness, as group-level equality requires treating equally qualified individuals differently based on group membership. As these measures all have their own advantages and disadvantages, fairness in algorithmic decision-making involves trade-offs (Pessach & Shmueli, 2023, p. 5). Thus, no singular definition can apply across all contexts.

2.2 Colorism and representation in digital media

Prominently, colorism can be defined as favoring Eurocentric facial features, such as lighter skin tones and straight hair, over Afrocentric features such as darker skin tones and textured hair types (Monk, 2021, p. 78). Skin tone stratification is a key aspect of this discriminatory practice and is likely tied to larger (historical) biases against Black or African-descended physical features. According to Hunter (2007, p. 238), racial discrimination can operate in two levels: race and color. The former of which entails that all

Black people, regardless of skin tone or physical appearance, will experience discrimination simply for being Black. The latter entails that the intensity and frequency of discrimination will differ based on skin tone. Both of these are a manifestation of the larger system of racism. Unlike racism, which creates social division between racial groups, colorism generates and maintains social division within racial groups (Bijou & Colen, 2022, p. 2).

Colorism historically emerged during European colonialism, during which it was built on a system that associated light skin with power, beauty, rationality, and civility, and dark skin with savagery, ugliness, and inferiority (Hunter, 2007, p. 238). Colorism is also deeply rooted in cultural norms, as many countries with European colonial histories (India, Philippines, Vietnam, etc.) remain to internalize and favor lighter skin tones and Anglo facial features, because European values were being enforced by the colonial regime (Hunter, 2007, p. 238). This is illustrated in Chattopadhyay (2019, p. 65)'s study, which highlighted the most common search algorithm on online dating site in India is a skin tone filter. More specifically, the usage of the word "fair", indicating the preference for lighter-skinned individuals. Similarly, scholars found that lighter-skinned Black individuals with "White" hair textures (e.g., straight or loosely curled hair) in Black communities were seen as prestigious and more socially desirable. (Monk, 2021, p. 78-79). This made it easier for these individuals to move up the social ladder. Moreover, skin tone discrimination still persists in South Africa even after its long history of colorism (Trammel, 2023, p. 63). Women in South Africa still partake in skin bleaching to achieve white skin. However, it is crucial to understand that context greatly influences whether race or skin color serves as the primary indicator of social status (Bijou & Colen, 2022, p. 2). For instance, in racially heterogeneous countries such as the United States, race becomes the signifier and main source of stigma. While racially homogenous countries such as those in the Caribbean (e.g., Jamaica) or Asia (e.g., India) skin color is the indicator of social status (Bijou & Colen, 2022, p. 2).

Colorism frequently leads to unequal opportunities and/or treatment depending on skin color (Lu et al., 2024, p. 812). This can impact health, employment, education, social interactions, and mental health issues. Edelman et al. (2017, p. 7) performed a field experiment on Airbnb and found that applicants and profiles from guests with distinctively African sounding names were less likely to be accepted compared to more White sounding names. Similarly, Stockstill and Carson (2022, p. 911) uncovered that although White employers rated both lighter- and darker-skinned Black applicants with distinctly Black names as equally employable, they nonetheless offered higher salaries to the lighter-skinned applicants. Interestingly, Childs (2022, p. 2) pointed out that light skin has always been associated with attractiveness and greater access to cultural capital and upward socioeconomic mobility. Studies show that people who are more attractive are seen as smarter and friendlier (Hunter, 2007, p. 241). Therefore, in job applications, lighter skin can unfairly benefit job applications, because attractiveness is a social construct influenced by racial aesthetics. Dark skin tones are continuously penalized in the job market in terms of earnings, unemployment rates, and even professional status (Hunter, 2007, p. 242). Moreover, Keyes et al. (2020, p. 692) found that colorism greatly affects the psychological well-being of dark-skinned individuals and leads to low self-esteem, internalized colorism, and harmful behaviors such as skin bleaching.

Digital representation often reflects skin tone biases. In May 2022, Google banned ads promoting skin-lightening products that implied one skin tone is superior. However, Lu et al. (2024, p. 815) found that this policy mostly targeted explicit messages that linked lighter skin to higher social status, beauty, or opportunities, while implicit colorist messages continued to appear, showing the limited impact of the update. Ads were promoting skin lightening for aesthetic reasons, thus subtly reinforcing colorist ideals. Similarly, retail advertising (either online or in person) has also been known for favoring lighter-skinned models over darker skin. Butkowski et al. (2022, p. 301) found that clothing brands such as Old Navy and Banana Republic favored lighter-skinned models on their website, despite their diverse cast. Particularly, the Banana Republic website appeared to display color-based tokenism, whereas one very dark-skinned model would pose amongst a generally lighter group. Skin lightening through photo editing is a common photographic technique that reinforces colorism (Butkowski et al., 2022, p. 290); this is categorized as dysconscious racism, an unconscious bias that associates lighter skin tones with beauty and professionalism (Butkowski et al., 2022, p. 303).

Colorism is also evident in popular culture such as in Latin American telenovelas where almost the entire cast is white except for maids who are light brown (Hunter, 2007, p. 240). In digital media, YouTube thumbnails with dark-skinned Black influencers were less likely to be clicked compared to lighter-skinned influencers (Love, 2022, p. 157). Interestingly, search engine ads are based on user engagement and commercial incentives. So, search engines reinforce colorism by ranking lighter-skinned individuals higher (Noble, 2013, p. 4). Additionally, Google images search results also provide harmful biases against Black people by linking images of Black women to stigmatization or hypersexualization (Noble, 2013, p. 4) or excluding them from search result for typing in certain keywords such as “beautiful women” or “professional hairstyles”. This invisibility and negative portrayals are forms of colorist algorithmic bias.

2.3 Curated feeds and perception of trust and fairness

To make their platforms more relevant and compelling, social media platforms rely on personalized algorithms (Eg et al., 2023, p. 1). This personalization occurs when online material adapts to the user's previous activities in an algorithmically created feedback loop (Bodle, 2019, p. 130). Algorithms filter and prioritize content for users based on demographics, online habits, and preferences. Some examples of this include featured recommendations based on previous purchases on Amazon.com, customized playlists on Spotify, Twitter trends, film recommendations on streaming platforms, and YouTube’s homepage recommendations. Users are presented with material that they have contributed to curating, thus they can benefit from the personalized web's convenience, effectiveness, interest, and relevancy (Bodle, 2019, p. 130). According to Lury and Day (2019, p. 22), there are two types of recommendation algorithms: collaborative filtering algorithms and content sharing algorithms. The former is based on behavior and preferences and will result in predictions based on what the user likes. It relies on patterns in user activity rather than demographic data. The latter recommends content based on features or characteristics of content you have already liked or interacted with.

However, while personalized suggestions are convenient, users still worry about who has access to their data and how it is being used (Head et al., 2020, p. 20; Yang et al., 2024, p. 3). This is due to algorithmic opacity, in which an algorithm lacks visibility of computational processes, and where users are unable to understand an algorithms' internal operations behind the recommendations that were provided (Paudyal & Wong, 2018, p. 193; Yang et al., 2024, p. 3). For example, on e-commerce platforms, users might discover item recommendations that are entirely unrelated to their previous purchases or past searches (Yang et al., 2024, p. 3). The complexity and lack of transparency makes it challenging to determine whether the algorithm complies with ethical requirements (Paudyal & Wong, 2018, p. 193). Eslami et al. (2019, p. 12) found that adding transparency can improve user interaction and increase trust in algorithmic systems. The transparency should include the existence of an algorithm in the platform, understanding how the system works without overwhelming the user, and understanding how the algorithm will affect them. Overall, trust will most likely increase.

Nonetheless, there is still uncertainty whether recommendation results accurately reflect user preferences or are reasonably accountable (Shin, 2020, p. 8). Woodruff et al. (2018, p. 6) offered some insights into perceptions of algorithmic fairness among marginalized users; they found that users do not like it when online platforms recommend personalized content based on their demographic characteristics (e.g., race, gender, ethnicity) and based on online behavior of other people with similar demographic characteristics. Users perceived such algorithmic personalization as not only unfair and intrusive, but also as perpetuating harmful stereotypes by assuming that individuals of the same demographic group share identical preferences and behaviors. Furthermore, participants in Woodruff et al. (2018, p. 6)'s study proposed three common causes of algorithmic unfairness: (1) a non-diverse population of programmers; (2) prejudiced online behavior by members of society; and (3) biased media representation. Therefore, they concluded that engagement and user trust will increase if algorithmic fairness is incorporated in products or online platforms. Riccio et al. (2024, p. 10) found that beauty filters on social media platforms such as Instagram and TikTok reinforce Eurocentric beauty standards and existing biases by lightening users' faces. As trust in algorithmic systems is closely related to users' perceptions of fairness, transparency, and representation, users may interpret personalized feeds that consistently favor lighter-skinned individuals as systemically racially biased and exclusionary. This bias perception will diminish the platforms' credibility and will decrease users' perceptions of its fairness or equity. Thus, exposure to algorithmic content that exhibits colorist bias will erode trust in the platform and perceptions of fairness.

H1*: Exposure to algorithmically biased curated feeds favoring lighter skin tones decreases platform trust*

H2*: Exposure to algorithmically biased curated feeds favoring lighter skin tones decreases perceived fairness*

Conversely, the presence of diversity (racially or gendered) in an algorithm produces positive expectation of platform trust and fairness (Chen & Sundar, 2024, p. 4; Wu et al., 2024, p. 7). Kordzadeh & Ghasemaghaei (2022, p. 394) highlights the stimulus-organism-response theory which contends that people's internal (psychological) moods are influenced by their surroundings, and this results in behavioral reactions. In this framework, algorithmic bias is the stimulus that influences user's perceptions (organism), which shapes behavioral responses such as trust. Chuan et al. (2024, p. 11) built their study off this model and found that participants who saw skin tone congruent recommendation perceived algorithms as fair and inclusive. However, those that were not recommended diverse skin tones reported lower fairness, which led to lower trust in the algorithm. Interestingly, they reported that individuals who have experienced colorism view it as even less fair and negatively affects perceived trust. Users are more likely to trust an algorithm when the recommendations align with their identity. Furthermore, the more familiar users are with an algorithm, the more inclined they are to trust it (Shin, 2020, p. 8). Cabiddu et al. (2022, p. 692) found that users' trust with an algorithm can develop over time; this is affected by familiarity with the system, social influence such as other users' shared experiences, and system-like characteristics such as reliability or consistency. Based on their findings, Chen and Sundar, (2024, p. 25) highlighted the importance of algorithmic transparency and pointed out that algorithms that exhibit racial diversity (either the labelers or the output) users are more likely to believe it to be fair and trustworthy.

H3: *Exposure to curated feeds with diverse skin tones increases platform trust*

H4: *Exposure to curated feeds with diverse skin tones increases perceived fairness of the platform*

2.4 Cross-group friendships as moderator

Friendships are healthy and important aspects of social life and serve many functions such as emotional-security, entertainment, support, intimacy, and many more (Fehr & Harasymchuk, 2022, p. 454; Bahns et al., 2015, p. 475). Friendships usually develop because people tend to be attracted to individuals that are similar to them; however, interacting with people that have different attitudes, values, beliefs, and experiences than our own can help us gain knowledge outside our own lived experiences (Bahns et al., 2015, p. 475). Known as cross-group friendships, these friendships are a type of interpersonal relationship between individuals from different racial, social, or cultural backgrounds (Page-Gould et al., 2022, p. 27). This is not the same as intergroup contact (e.g., a quick conversation with a person of a different race) as it is a close, ongoing, and emotionally meaningful relationship with an individual from a different social group.

Bahns et al. (2015, p. 484) found that people who value diversity are more likely to have diverse social networks. This is primarily demonstrated by elements like inclusive attitudes or involvement in multicultural environments, which entail acknowledging and appreciating the diversity of individuals from various backgrounds and cultures (Bahns et al., 2015, p. 477). Similarly, Davies et al. (2011, p. 342) highlighted that people might be motivated to form cross-group friendships to achieve social goals such as intimacy, social approval, or self-expansion. However, by sharing personal stories, cross-group friendships can develop emotional closeness and reduce anxiety. As these relationships deepen, trust and empathy will enhance by exchanging shared experiences and perspectives (Davies et al., 2011, p. 342). Overall, this illustrates that having positive relationships with individuals in marginalized groups offers a positive view of the marginalized individual while decreasing perceptions of prejudice and racial inequalities (Dixon et al., 2010, p. 77). This suggests that cross-group friendships can lead to greater awareness of societal issues such as systemic racism and colorism.

Unfortunately, there appears to be little to no research that directly examines how the diversity of one's social circle might influence perceptions of social media platform fairness and trust when encountering colorist algorithmic biases; but there are substantial studies about how users' experiences with diversity and/or colorism could influence their perceptions of platform trust and fairness. For instance, Chen and Sundar (2024, p. 25), Mourali et al. (2025, p. 15), and Chuan et al. (2024, p. 11) showcased the importance of diversity in AI systems and training data, and that historically disadvantaged groups are more sensitive to algorithmic bias, which likely stems from experiences with discrimination. With the large body of evidence that showcases the impact of colorism on media representation along with evidence that supports that cross-group friendships reduce prejudice beliefs, it is safe to say that users that are open to maintaining cross-group friendships are able to understand, recognize, and be aware of certain (racial) biases, including algorithmic fairness in digital spaces. While a decrease in platform trust is expected following exposure to algorithmically biased curated feeds favoring lighter skin tones, this effect is anticipated to be stronger among individuals with more cross-group friendships, as they are more likely to recognizing racial bias.

H5: *Cross-group friendships moderate the relationship between exposure to colorist algorithmically biased curated feeds and (a) platform trust and (b) perceived fairness*

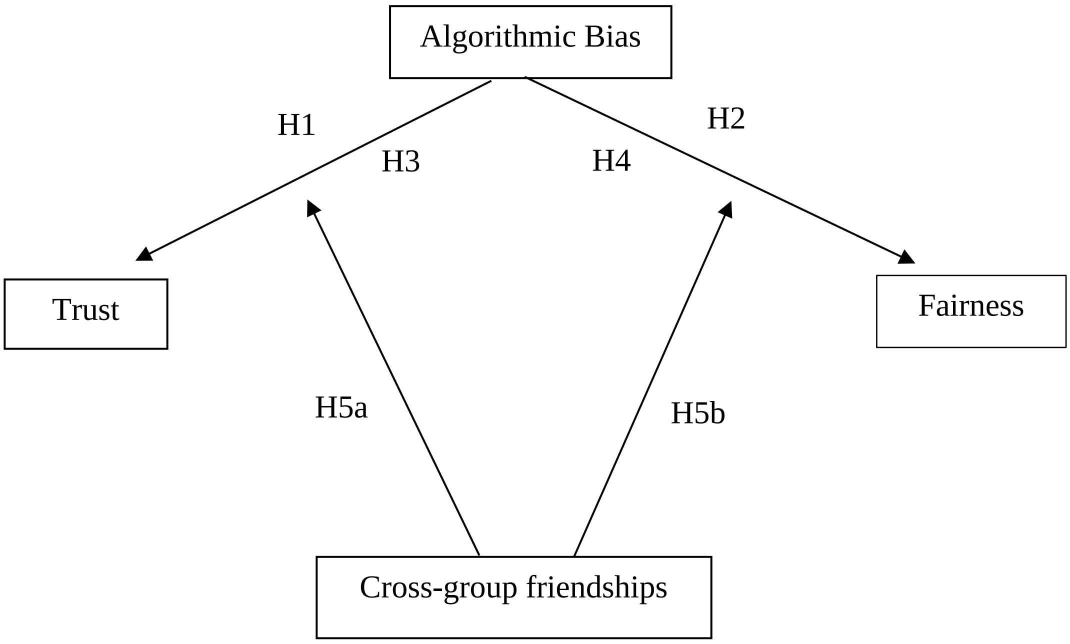


Figure 1. *Conceptual model of hypotheses*

3. Methodology

3.1 Quantitative approach

3.1.1 Research design

Since the aim of this thesis is to examine how colorist algorithmic biases can impact user's perception of platform trust and fairness, a quantitative survey experiment will be utilized. Social science experiments are useful for testing causal relationships by manipulating one variable and comparing the outcomes (Neuman, 2014, p. 283). Creswell & Creswell (2018, p.) point out that quantitative studies use hypotheses to predict expected outcomes of relationships among variables. These are used for deductive theorizing, as the hypotheses are formulated based on existing theories and literature and are then tested using data (Neuman, 2014, p. 70). In this study, the hypotheses are formulated based on prior research on algorithmic bias and colorism in digital spaces and are tested through controlled exposure to different social media feeds to examine potential changes in perceptions of trust and fairness.

This study will present participants with curated mockup feeds from Instagram to examine how various skin tone compositions affect trust and fairness perception of platforms. Additionally, it allows for controlled exposure to different curated feeds while reducing potential social desirability bias. Due to ethical and practical constraints, many social science questions cannot be studied. However, a survey experiment can simulate interventions or conditions without invasive real-world interventions or raising moral concerns (Neuman, 2014, p. 283). This method is particularly appropriate for this study because conducting a real-life intervention through users' actual social media feeds would be ethically complex and technically challenging. By simulating this through mock Instagram feeds, the survey experiment enables controlled exposure while preserving the participant's privacy and well-being. Furthermore, survey experiments are suitable for micro-level—studying individual psychological effects or small-group phenomena—instead of large-scale societal trends (Neuman, 2014, p. 283). Therefore, this allows the study to isolate the impact of different algorithmic conditions on user's perceptions and assess the moderating role of friend group diversity. Lastly, survey experiments are efficient (can be conducted online to reach a wide diverse sample) and could enhance a diverse and scalable participant pool, as well as the replicable structure of the research design will ensure that future research can build upon these findings.

3.1.2 Experimental design

This study employed a between-subject experimental design, as each participant will be exposed to one of the two conditions: a mock social media feed consisting of predominantly white individuals or a feed with a racially diverse representation of skin tones. The goal is to examine whether the racial diversity (or lack thereof) in algorithmically recommended content influences users' perceptions of trust and fairness of social media platforms. The study adopted a pre-test and post-test design to assess changes in attitudes before and after exposure to the feeds. This design also included a moderator variable of the level of racial diversity in participant's social networks.

3.2 Sampling

The experiment was conducted in an online environment using Qualtrics. This was done in an effort to easily reach participants from different countries in a convenient and timely manner and allow them to retain their anonymity (Neuman, 2014, p. 283). A non-probability sampling method was employed, where participants were selected based on their active use of social media platforms. This sampling method, known as purposive sampling, guarantees that the sample consists of individuals with a specific characteristic (Neuman, 2014, p. 273), and could meaningfully engage with the visual content presented in the questionnaire. To achieve this, the questionnaire was sent out and frequently reposted on various social media platforms and groups such as Instagram, WhatsApp, Facebook, and Reddit. Additionally, snowball sampling was also used by asking people to share the questionnaire with others. According to Babbie (2017, p. 197), "snowball" refers to the process of gathering participants as each participant suggests another participant.

The data was collected in a span of a week, from April 29th, 2025, to May 6th, 2025, and resulted in 281 participants. Participants that took part in the pilot test were removed, as well as dropouts. After data cleaning, N = 216 were used for further analysis, with 105 assigned to the condition with lighter skin tones and 111 to the condition with diverse skin tones. The sample consisted of 166 females (76.9%), 38 males (17.6%), 9 individuals who identified as "non-binary" (4.2%), and 3 individuals that marked "prefer not to say" (1.4%). The age range of the participants was 19 –70 years old, with the most frequently mentioned ages being 22 (19.2%), 23 (11.2%), and 21 (8.4%). The participants nationalities originated from 40 countries, with the most common ones being Suriname (42.6%), The Netherlands (13.9%), and France (5.1%). As for education level, the recurring highest education level was bachelor's degree (50.5%), followed by master's degree

(20.1%), and a high school diploma (13.6%).

3.3 Procedure

Upon clicking the survey link, participants were presented with an introductory message that briefly explains the aim of this study and assuring them anonymity and confidentiality. Before participants could start with the survey, informed consent was obtained by having them confirm that they are 18 years or older and agree to participate. The questionnaire began with a pre-test in which the dependant variable will be measured before the treatment (Neuman, 2014, p. 291). Social media use (what social media platform(s) they use and what kind of content they consume) and perceptions of trustworthiness and fairness were measured.

Then, participants were exposed to the experimental manipulation of four mock Instagram feeds (created using Picsart) consisting of predominantly white individuals or feeds with racially diverse skin tones. Participants were randomly assigned to one of the two conditions. After viewing the feeds, participants completed a post-test in which they answered the same trust and fairness items again to detect and measure changes in their perceptions (the dependant variable) (Neuman, 2014, p. 291). A manipulation check followed, which asked participants whether they noticed diversity in the feeds they viewed. To measure the moderator, participants were asked how many people in their social circle have a different ethnic background than them, followed by statements about their openness and comfort with racially diverse people. The final section of the questionnaire consisted of demographic questions such as age, gender, education level, and nationality. The questionnaire concluded with a debrief, informing participants that the feeds they were shown were fake and were made solely for the purpose of this study.

3.4 Operationalization

3.4.1 Stimulus material

The mock Instagram feeds used in the experimental manipulation were designed to resemble real personalized Instagram content on the “explore page”, consisting of a 3x3 grid of images and video thumbnails. Each feed differed in curated content and consisted of makeup, fashion, food, fitness, and lifestyle related content. In the predominantly white condition, all individuals shown in the feeds had light skin tones and Eurocentric features. In the diverse condition, individuals of various racial and ethnic backgrounds were shown. The feeds were standardized in layout and number of posts. All the visual elements were carefully balanced in terms of image type and aesthetic to ensure that racial representation was the only meaningful variable that differed between the conditions (see Appendix A).

3.4.2 Validity and reliability

Before testing the variables, the validity and reliability of the measurements that were used were tested. Validity and reliability are crucial parts of social science research because they are ideas that help us determine the truthfulness, credibility, and believability of the findings (Neuman, 2014, p. 291) Put simply, validity refers to how accurately an existing measurement can measure what it is supposed to measure (Creswell & Creswell, 2018, p. 200). A handful of scales and variables were included to effectively measure the research topic at hand. To ensure this, content validity was employed. This refers to the extent to which the items of a scale accurately represent all aspects of the construct being measured (Neuman, 2014, p. 291, 216). Content validity was guaranteed in this study by adapting measurements from previously validated scales related to algorithmic fairness and trust. Furthermore, a pilot test was conducted with a small group of participants. Based on their feedback, some items were re-worded for better clarity and align better with the context of the study. For example, all items were modified to explicitly refer to Instagram's algorithm, ensuring that participants understood the constructs in relation to the platform being studied.

Furthermore, construct validity—which refers to a measurement validity that assesses whether the items in a scale accurately reflect the concepts it is meant to measure (Neuman, 2014, p. 217; Creswell & Creswell, 2018, p. 200)—was also checked. According to Neuman (2014, p. 217), construct validity is supported if the items on a scale operate consistently. Therefore, to ensure construct validity, one item of the scales "Perceived fairness" and "Comfort with differences" were removed as they did not clearly align with the intended theoretical focus of perceived fairness and cross-group friendships.

To ensure the reliability of the measurements, a reliability analysis was performed. Reliability refers to the consistency and repeatability of a scale (Creswell & Creswell, 2018, p. 200). A reliability analysis establishes the internal consistency— the degree to which the scale items measure the same underlying concept— of the study (Creswell & Creswell, 2018, p. 200). These were evaluated with the Cronbach's α. An alpha of .65-.80 is often considered as adequate reliability (Vaske et al., 2016, p. 165). An alpha above .80 has very good reliability. The Cronbach's α was calculated for each adopted scale to measure their internal consistency (see next paragraph).

3.4.3 Measurements

Perceived trust. Measurement items of perceived trust were adopted from Shin (2021, p. 9). The scale consists of 10 subscales, however, the current study made use of the subscale "Trust" ($\alpha = .87$, $M = 11.75$, $SD = 3.73$). The subscale was assessed via three items and consisted of the following items: "I trust the recommendations by algorithms-driven services", "Recommended items through algorithmic processes are trustworthy", and "I believe that the algorithm service results are reliable". To match the essence of the current study, the items were modified to the following, "I trust the recommendations by Instagram's algorithm", "Recommended content through Instagram's algorithm are trustworthy", and "I believe that Instagram's algorithmic feed results are reliable". Each item was measured on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). The items were entered into a confirmatory factor analysis using Principal Components extraction with Direct Oblimin rotation, $KMO = .74$, $\chi 2\ (3) = 319.38$, $p < .001$. The resulting model consisted of one factor and explained 79.5% of the variance in platform trust. With only one component having an Eigenvalue above one (Eigenvalue of 2.38), the analysis shows that the three items fall under a one-dimensional scale. The factor loadings and Cronbach alphas of each factor are presented in Table 1.1.

Perceived fairness. To measure perceived fairness, the same scale by Shin (2021, p. 9) was borrowed. The subscale titled "Fairness" was utilized and consisted of three items: "The system has no favoritism and does not discriminate against people", "The source of data throughout an algorithm and its data sources should be identified, logged, and benchmarked", and "I believe the system follows due process of impartiality with no prejudice". The items were measured on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Before conducting the main study, a pilot test of the questionnaire was performed, which resulted in the modification of the items for better transparency and understanding. Similarly, the items were also modified to match the essence of the current study. However, internal consistency was revealed to be extremely low ($\alpha = .15$, $M = 12.22$, $SD = 2.74$) Additionally, after entering the items in a confirmatory factor analysis using Principal Components extraction with Direct Oblimin rotation, it was revealed to have a low Kaiser-Meyer-Olkin value of .56. This indicates that the items do not correlate well with each other. Moreover, the Bartlett's Test of Sphericity was significant, $\chi 2(3) = 65.50$, $p < .001$. The resultant model explained 53.3% of the variance in algorithmic fairness, with only one component having an Eigenvalue above one (Eigenvalue of 1.60).

After removing the item "I believe Instagram should be transparent about the data it collects for its recommendations, document that data properly, and regularly evaluate its fairness and accuracy" internal consistency improved drastically (α = .82, *M* = 6.30, *SD* = 3.05). This is likely due to the fact that the item was long and multifaceted, which may have confused respondents or caused inconsistent interpretations compared to the more straightforward items. The factor loadings and Cronbach alphas of each factor are presented in Table 1.2.

Cross-group friendships. To measure the moderation variable of cross-group friendships, a five-item subscale "Comfort with differences" adapted by Fu et al. (2018, p. 133)—which measure the comfort and willingness to form cross-group friendships—was used. One example item is "It's really hard for me to feel close to a person from another race". The items were measured on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). All items were reverse coded; therefore, higher scores indicate greater comfort/openness with racial diversity in friendships, whereas lower scores reflect greater discomfort. The items were entered into a confirmatory factor analysis using Principal Components extraction with Direct Oblimin rotation, *KMO* = .69, $\chi 2$ (10) = 249.42, $p <$ .001. The resulting model consisted of two factors and explained 69.2% of the variance. The first factor included four items about racial avoidance or discomfort, which explains 48.7% of the variance. The second factor loaded only one strong item about value alignment, which explained 20.5% of the variance. The item "It is very important that a friend agrees with me on most issues" was removed due to poor alignment with the rest of the items and low item correlation. The remaining four items resulted in acceptable internal consistency (α = .76, *M* = 6.08, *SD* = 3.19). The factor loadings and Cronbach alphas of each factor are presented in Table 1.3.

Table 1.1.

Factor loadings, explained variance, and reliability of the factor found in the scale "Perceived trust"

Item	*Platform trust*
I trust the recommendations by Instagram's algorithm.	.88
Recommended content through Instagram's algorithm are trustworthy.	.89
I believe that Instagram's algorithmic feed results are reliable.	.90
R^2	.79

Cronbach's α	.87

Table 1.2.

Factor loadings, explained variance, and reliability of the factor found in the scale "Perceived fairness"

Item	*Algorithmic fairness*
Instagram's algorithm shows no favoritism and does not discriminate against people.	.82
I believe Instagram's algorithm follows due process of impartiality with no prejudice.	.81
R^2	.53
Cronbach's α	.82

Table 1.3.

Factor loadings, explained variances, and reliability of the two factors found in the scale "Comfort with differences"

Item	*Racial discomfort*	*Value alignment*
I am only at ease with people of my own race.	.86	
It's really hard for me to feel close to a person from another race.	.84	
I often feel irritated by persons of a different race.	.72	
Getting to know someone of another race is generally an uncomfortable experience for me.	.64	
It is very important that a friend agrees with me on most issues.		.92
R^2	.49	.21
Cronbach's α	.76	

3.5 Data analysis

IBM SPSS 27 was used to analyze the data. To evaluate the first (H1*: Exposure to* *algorithmically biased curated feeds favoring lighter skin tones decreases platform trust*), second (H2: *Exposure to algorithmically biased curated feeds favoring lighter skin tones decreases perceived fairness*), third (H3: *Exposure to curated feeds with diverse skin tones increases platform trust*) and fourth hypothesis (H4: *Exposure to curated feeds with diverse skin tones increases perceived fairness of the platform*) a paired sample t-test was performed to analyze the perceptions of trust and fairness before and after exposure to the experimental conditions. This analysis was used to compare the responses from the participants before and after exposure to the curated Instagram feeds. This is particularly suitable for assessing whether the experimental manipulation successfully produced a statistically significant change in participants' perceptions of trust and fairness. To measure the moderation effect of cross-group friendships (H5: *Cross-group friendships moderate the relationship between exposure to colorist algorithmically biased curated feeds and (a) platform trust and (b) perceived fairness*), a moderation analysis was performed using PROCESS Model 1, an extension that was downloaded as an additional function to the software. This extension was used to simplify the moderation analysis by automatically generating a clear statistical output.

4. Results

4.1 Descriptive analysis of the pre- and post-test scores across demographics

To provide further insights into the distribution of responses, the mean scores of the pre- and post-test scores for perceived trust and fairness were calculated across the key demographics. The results revealed that Instagram is the most used social media platform among the participants (81.5%), followed by TikTok (59.3%), and Facebook (44.4%). The type of content that the participants consumed was very diverse, but the most common ones were cooking/food content (67.6%), fashion (57.9%), and lifestyle content (55.6%). It is important to note that participants were given the option to select more than one platform and type of content.

When looking at age groups in the lighter skin tones condition, participants aged 19-25 had a slight increase or stability in trust after exposure to the feeds, while there is a noticeable decrease in fairness after exposure to the feeds. Older participants between 32-70 had overall more neutral or a slight decrease in trust after exposure to the feeds, as is the same with fairness. This may suggest that younger participants have more trust in platforms like Instagram. This is likely due to more frequent use or familiarity with the platform. In the diverse skin tones condition, participants aged between 19-26 had for the most part an increase in trust after exposure to the feeds, whereas participants aged between 28-68 had somewhat polarizing trust perceptions but did increase for the most part. Fairness perceptions remained largely neutral across all age groups (see Table B1 and B2 in Appendix B).

For gender, female participants (N = 78) in the lighter skin tone condition showed a neutral or slight decrease in trust from M = 3.98 (SD = 1.22) before exposure to M = 3.80 (SD = 1.31) after exposure to the feeds. Male participants (N = 21) in this condition also have a neutral or slight decrease in trust after exposure from M = 3.52 (SD = 1.35) to M = 3.46 (SD = 1.41). The same can be said for the non-binary participants (N = 5) that went from M = 3.27 (SD = .92) to M = 3.07 (SD = .80). This indicates that there is a relatively stable perception in this condition across all genders. Fairness perceptions among male and non-binary participants were relatively the same as the trust perceptions. However, female participants had a higher decrease in fairness perceptions after exposure to the feeds from M = 3.08 (SD = 1.28) to M = 2.88 (SD = 1.41). Furthermore, mean scores in the diverse skin tones condition were overall stable across all gender groups (see Table B1 and B2 in Appendix B).

Across nationalities, the largest group (participants form Suriname) (*N* = 50) in the lighter skin tone condition had stable perceptions in trust and fairness after exposure to the feeds. However, the second largest group (the Netherlands) (*N* = 14), had a slight decrease in trust perceptions after exposure to the feeds from *M* = 3.55 (*SD* = 1.00) to *M* = 3.19 (*SD* = 1.48), but stable perceptions in fairness. Although these scores are quite stable, it does indicate the differing levels of sensitivity to racial representation in algorithmic systems across cultural contexts. Interestingly, there was a noticeably larger decrease in both trust and fairness perceptions after exposure to the feeds in the third largest nationality group (France) (*N* = 4). Perceptions of trust and fairness remained stable or slightly increased in the diverse skin tones condition across all nationality groups (see Table B1 and B2 in Appendix B).

The descriptive results of cross-group friendships show that almost half of the participants (49.5%) scored the highest score (7.00). This suggests that the majority of the participants show very high openness to individuals of different races. Furthermore, 12.5% of the participants scored 6.75 and 6.9% of the participants scored 6.50. This further indicates that there is a skew toward higher openness to cross-group friendships. Very few participants scored below 4.00 with only two participants scoring below 3.00. Overall, these results suggest that the sample generally expresses positive attitudes towards racially diverse friendships.

4.2 Effects of curated feeds on trust and fairness

This next part of the results section will focus on hypothesis testing. To analyze the effects of algorithmically curated Instagram feeds on users' perceptions of trust and fairness, a series of paired sample t-tests were conducted. For participants exposed to feeds with predominantly lighter skin tones with trust as the dependent variable results indicated that there was no significant difference between platform trust before exposure to the feeds (M = 3.86, SD = 1.24) and after exposure to the feeds (*M* = 3.71, *SD* = 1.34), *t*(104) = 1.66, *p* = .101. Similarly, there was no significant difference between platform fairness before exposure to the feeds (*M* = 3.03, *SD* = 1.28) and after exposure to the feeds (*M* = 2.93, *SD* = 1.46), *t*(104) = 1.32, *p* = .192. There was no significant change in perceptions of platform trust and fairness after exposure to the Instagram feeds with predominantly white individuals. H1 and H2 are rejected.

Conversely, participants exposed to feeds with racially diverse skin tones had a significant increase in platform trust before exposure to the feeds (*M* = 3.97, *SD* = 1.25) and after exposure to the feeds (*M* = 4.14, *SD* = 1.35), *t*(110) = 2.12, *p* = .037, as well as a significant increase in perceived fairness before exposure to the feeds (*M* = 3.18, *SD* = 1.43) and after exposure to the feeds (*M* = 3.35, *SD* = 1.56), *t*(110) = 2.08, *p* = .040. Therefore, participants that were exposed to the feeds with diverse skin tones viewed Instagram as more trustworthy and fairer. H3 and H4 are accepted.

4.3 The moderating role of cross-group friendships

A moderation analysis was conducted using SPSS PROCESS Model 1 to test whether cross-group friendships moderate the role between algorithmically biased curated feeds and perceived trust. Hence, it is testing whether effects differ between individuals who have higher scores of comfort with racial diversity in their friendships and those that have lower scores of comfort. The overall model revealed to be significant, $F(3, 212) = 2.91$, $p = .035$, $R^2 = .04$. However, the interaction between algorithmically curated feeds and cross-group friendships was not significant ($\beta = .27$, $SE = 0.23$, $t = 1.18$, $p = .238$). This means that friend group diversity does not statistically moderate the relationship between algorithmically biased curated feeds and perceived trust. H5a is rejected. Table 2.1 presents an overview of the moderation analysis outcome.

Table 2.1

Moderation analysis coefficients table with trust as criterium

	B	SE	t	p	95% CI
Constant	3.71	0.13	28.45	.000	[3.4543, 3.9686]
Algorithmically biased curated feeds	0.45	0.18	2.48	0.14	[0.0917, 0.8095]
Cross-group friendships	0.03	0.15	0.22	.823	[-0.2684, 0.3369]
Interaction	0.27	0.23	1.18	.238	[-0.1809, 0.7249]

Note. B = Unstandardized coefficients; SE = Standard error; CI = Confidence Interval; DV= Trust after exposure

As for whether cross-group friendships moderate the role between algorithmically biased curated feeds and perceived fairness, the overall model revealed to be significant, $F(3, 212) = 2.93$, $p = .034$, $R^2 = .04$. However, the interaction between algorithmically curated feeds and cross-group friendships was not significant ($\beta = .18$, $SE = 0.26$, $t = 0.71$ $p = .483$). This means that friend group diversity does not statistically moderate the relationship between algorithmically biased curated feeds and perceived platform fairness.

H5b is rejected. Table 2.2 presents an overview of the moderation analysis outcome.

Table 2.2

Moderation analysis coefficients table with fairness as criterium

	B	SE	t	p	95% CI
Constant	2.92	0.15	19.79	.000	[2.6277, 3.2090]
Algorithmically biased curated feeds	0.46	0.21	2.44	0.26	[0.0917, 0.8095]
Cross-group friends	0.18	0.17	1.04	.298	[-0.1611, 0.5231]
Interaction	0.18	0.26	0.70	.483	[-0.3293, 0.6943]

Note. B = Unstandardized coefficients; SE = Standard error; CI = Confidence Interval; DV= Fairness after exposure

4.4 Comparison of post-test trust and fairness scores between the conditions

After analyzing how trust and fairness perceptions changed before and after exposure to the stimulus, it is also important to compare the post-exposure scores across the conditions. To examine this, an independent sample t-rest was conducted because a paired sample t-test cannot compare between conditions. The results indicated that there was a significant difference in perceived trust scores for participants exposed to feeds with lighter skin tones (M = 3.71, SD = 1.34) and those exposed to the feeds with diverse skin tones (M = 4.14, SD = 1.35), $t(214) = 2.33$, $p = .021$. Similarly, fairness scores were significantly higher in the diverse skin tones condition (M = 3.35, SD = 1.56) than in the lighter skin tones condition (M = 2.93, SD = 1.46), $t(214) = 2.03$, $p = .044$.

4.5 Significance of manipulation check

A manipulation check was conducted to assess whether participants noticed the diversity of skin tones in the Instagram feeds they were shown. A cross-tabulation revealed that 92.8% of those in the diverse skin tones condition correctly noticed the diversity, while 40% of those in the lighter skin condition incorrectly noticed the diversity. Furthermore, a chi-square test of independence revealed a statistically significant relationship between the conditions participants were assigned to and whether they reported noticing diverse skin tones in the feeds, X^2 (1, N = 216) = 68.15, $p < .001$. These results suggest that the manipulation was successful in generating different perceptions of diversity across the conditions.

Table 2.3

Hypotheses results

Hypothesis	**Result**
H1: Exposure to algorithmically biased curated feeds favoring lighter skin tones decreases platform trust	Rejected
H2: Exposure to algorithmically biased curated feeds favoring lighter skin tones decreases perceived fairness	Rejected
H3: Exposure to curated feeds with diverse skin tones increases platform trust	Accepted
H4: Exposure to curated feeds with diverse skin tones increases perceived fairness of the platform	Accepted
H5a: Cross-group friendships moderates the relationship between exposure to colorist algorithmically biased curated feeds and platform trust	Rejected
H5b: Cross-group friendships moderates the relationship between exposure to colorist algorithmically biased curated feeds and perceived fairness	Rejected

5. Conclusion

This thesis aimed to examine how colorist algorithmic bias affects users' perceptions of trust and fairness on Instagram. In this day and age, algorithms shape the content people consume daily, and it is vital to understand whether these systems reinforce societal biases and how people respond to them. Additionally, this thesis also explored how racial diversity in people's friend groups might moderate these perceptions.

5.1 Main findings

The goal of this study was to examine how curated algorithmically biased Instagram feeds impact perceptions of perceived trust and platform fairness. Additionally, this study also examined the moderating role of friend group diversity. A between-subject experiment was conducted in which participants were randomly assigned a condition with social media feeds with either predominantly lighter skin tones or diverse skin tones. As seen in Table 2.1, two out of the six hypotheses posed were accepted. The first hypothesis explored the relationship between trust and Instagram feeds with predominantly white individuals. The results showed that there was no significant change in perceptions of trust before and after exposure to the feeds.

Similar results were found for the second hypothesis, in which there were no significant changes in perceptions of platform fairness. The third and fourth hypothesis were accepted, which revealed that there was a significant change in perceptions of perceived trust and perceived platform fairness after exposure to social media feeds with diverse skin tones. Furthermore, cross-group friendships were tested as a moderator, but were not found to be a factor in strengthening the relationship between algorithmically biased feeds and perceived trust and platform fairness. Lastly, the results also revealed that the manipulation check did have a significant effect on participants noticing diversity in the feeds. Overall, the model revealed that there is a significant difference in perceptions of trust and fairness when participants are shown feeds with predominantly lighter skin tones versus diverse skin tones.

5.2 Discussion

The result revealed that there is a significant difference in perceptions of trust and fairness scores for participants exposed to the different feeds. Furthermore, these findings also show that feeds consisting of predominantly lighter skin tones versus diverse skin tones can greatly influence how much someone perceives a platform as trustworthy or fair. This can also be understood through the Stimulus-Organism-Response (S-O-R) theory

(Kordzadeh & Ghasemaghaei, 2022, p. 394), which proposes that individual's psychological states (organism) are influenced by external stimuli, in this case, racial diversity in algorithms. When users are exposed to inclusive content (stimulus), it positively affects their perceptions of fairness and trust, which will shape their behavioral response. Hence, this suggests that visible diversity in algorithmically curated content can create more positive user experiences, leading to increased engagement and user trust, which are crucial for boosting perceptions of fairness in digital environments. (Woodruff et al., 2018, p. 6). Overall, this finding provides an answer to the research question *"To what extent do colorist algorithmic biases impact users' perceived trust of and platform fairness of curated social media feeds?"*

When it came to Instagram feeds consisting of predominantly white individuals, it was found that exposure to these feeds did not decrease perceptions of trust and fairness among the participants. This finding contradicts prior research by Woodruff et al. (2018, p. 6) and Eslami et al. (2019, p. 12), which suggested that trust is most likely to increase if algorithmic transparency (understanding how the algorithm works and how it will affect the user) and fairness are present. In Woodruff et al. (2018, p. 6)'s study, they explored through interviews users' perceptions of algorithmic bias when it was explicitly linked to demographic profiling or stereotyping. This thesis utilized an experiment to test reactions to visual diversity cues in curated content. It is possible that such dominant or hegemonic representations (lighter skin tones in curated content) are perceived as normal and do not encourage critical reflection. Alternatively, exposure to more diverse representations might have activated more recognition of inclusivity, which in turn enhances trust and fairness perceptions. This suggests that perceptions are likely to be shaped by the presence of inclusive content rather than the absence of it.

Additionally, studies have shown that lighter skin is often associated with attractiveness and attractive individuals are often perceived as more friendly and trustworthy (Childs, 2022, p. 2; Hunter, 2007, p. 24; Stockstill & Carson, 2022, p. 911). This could also explain why participants did not perceive the lighter-skinned Instagram feeds as unfair or untrustworthy and may have implicitly influenced their judgments. Algorithmic bias favoring lighter skin tones may not have been consciously acknowledged as problematic, especially among participants that do not actively reflect on or experience colorism. Previous studies have shown that prior experiences with colorism can influence perceptions of algorithmic bias. This could also explain the general neutral stance participants have, as illustrated in the pre- and post-test mean scores for trust and fairness (approximately 3.5-4.1). Participants likely do not hold strong opinions about the trustworthiness and fairness of algorithmic systems. Nevertheless, future research could build on these findings by using quantitative methods to directly measure participants' awareness or lived experiences with colorism and test how this affects their recognition of algorithmic bias.

Furthermore, in line with Cabiddu et al. (2022, p. 692), it is possible that the lack of familiarity with Instagram's algorithm or lack of awareness about how such systems operate could have influenced how participants evaluated trust and fairness. When users are unaware of how algorithmic curation functions or how it affects them, they might not recognize subtle differences in representational diversity as meaningful, even if those designs are meant to increase inclusivity. Consequently, even though visible diversity was present in the feeds, it might not have been consciously perceived as significant enough to influence their beliefs of trust and fairness.

As expected, participants' trust and platform fairness did increase after being exposed to the feeds with diverse skin tones. This aligns with previous works that highlight the importance of diversity in algorithmic systems and recommended content (Chuan et al., 2024, p. 11; Chen & Sundar, 2024, p. 4; Mourali et al., 2025, p. 15; Wu et al., 2024, p. 7). These findings demonstrate that when participants are presented with racially inclusive content, they are more likely to view the platform as unbiased and trustworthy. However, the finding could also align with Shin (2020, p. 8)'s argument that an algorithm is viewed as trustworthy when the recommended content aligns with the users' identity. It is possible that participants found the content on the feed to be relatable, thereby perceiving it as trustworthy.

A surprising outcome of this study is that friend group diversity was not a significant moderator. As mentioned in section 2.4, there are no pre-existing studies that directly research how the diversity of one's social circle might influence perceptions of social media platform fairness and trust when encountering colorist algorithmic biases. This study offers insights into this by confirming that people might still recognize colorist algorithmic biases regardless of the diversity of their social networks. A possible explanation for this could be that participants' awareness of bias was not significantly shaped by their interpersonal experiences. This partially contradicts Bahns et al. (2015, p. 475), who argue that interacting with individuals who have different attitudes, values, beliefs, and experiences can increase one's understanding beyond their own lived experiences.

Another possible explanation for this could be the lack of variety in distribution scores. As mentioned in section 4.1, almost half of the participants gave similar answers; this resulted in a lack of variety that may have limited the ability to detect a potential moderation effect. There was nothing to compare between the high and low values, hence why the model could not detect an influence in the sample. This lack of variation could be attributed to the sampling method. Since the study used snowball sampling and relied heavily on social networks, it is likely that many participants shared similar views on comfort and openness to cross-group friendships. This homogeneity ultimately resulted in reducing the effectiveness of the moderation analysis. A more representative or stratified sampling method could have broadened the range of scores and perspectives and offered a potential moderating effect.

Ultimately, the findings of this thesis offer several practical implications. For starters, it is clear that social media platforms could benefit from more racially inclusive content. Platform developers could promote greater user trust and fairness by designing more inclusive algorithms. To avoid reinforcing harmful biases, developers should ensure that recommendation systems are trained on diverse datasets. Additionally, the findings also reinforce the need for transparency in how content is curated, especially on platforms that use personalized algorithms. Lastly, these findings encourage critical awareness of how algorithms influence digital experiences for minority groups.

5.3 Theoretical implications

This thesis extends the literature on algorithmic bias and fairness by providing academic evidence that racial diversity in algorithmically curated feeds on social media enhance users' perceptions of platform trustworthiness and fairness. It contributes to the growing body of research that challenges the notion that decisions made by automated algorithms are inherently fair or objective (Pessach & Shmueli, 2023, p. 1; Jacoba et al., 2023, p. 433). It adds depth to the definitions of algorithmic bias and fairness and builds on theoretical models like the Stimulus-Organism-Response (S-O-R) theory by offering insights into how psychological states can be influenced through visual representation within algorithmically curated content (Kordzadeh & Ghasemaghaei, 2022, p. 394). This thesis also contributes to methodological theory-building in algorithmic fairness research by utilizing an experimental design with visual stimuli. Previous studies relied on qualitative approaches through self-reported attitudes or interviews, whereas this study operationalizes colorism and algorithmic fairness in a quantifiable way.

In addition, it highlights the unique role colorism can play in algorithmic bias and fairness, offering valuable insights and a more nuanced theoretical perspective into how colorist preferences can be perpetuated through recommendation systems on platforms such as Instagram (Jarvis & Quinlan, 2022, p. 137; Trammel, 2023, p. 60; Ryan-Mosley, 2021, par. 12). Importantly, this thesis introduces colorism as a specific lens rather than race as a broad category. This is important in understanding racial bias and its digital manifestations and contributes to the growing understanding of how algorithmic systems replicate not just racism but specifically colorist hierarchies.

Furthermore, the findings of this thesis demonstrate that dominant or hegemonic forms of representation may be perceived as neutral by users. This intersects with theories of visual media and hegemonic normativity in digital spaces, where the “default” (lighter skin tones) is rarely questioned unless there is a clear presence of diversity. This highlights the need for future theoretical models to incorporate factors such as user awareness, critical media literacy, and algorithmic transparency as mediators for perceiving algorithmic bias.

This research also introduced the exploratory moderation variable of friend group diversity to examine whether interpersonal exposure to racial diversity (or the lack of) might shape users’ attitudes towards colorist algorithmic bias. Considering the outcome of the moderation analysis, this study provides the interesting theoretical insight that users might recognize algorithmic fairness regardless of the racial diversity in their friend group(s). Theoretically, this challenges the assumption rooted in diversity theories that interpersonal diversity enhances awareness of social inequalities. This suggests a potential for alternative moderation variables when researching algorithmic bias such as experiences with discrimination, media literacy, alignment with recommended content, user identity, or ideological orientation.

5.4 Limitations and suggestions for future research

Like many works, this thesis encountered a handful of limitations that should be acknowledged for future studies. The first limitation emphasizes the ecological validity and generalizability of the findings. This entails that the artificial nature of the experiment may not accurately reflect real life behavior from the participants (Neuman, 2014, p. 468). Even though the mock Instagram feeds were designed with the intent to identically resemble personalized feeds, participants might still respond differently from the controlled setting versus organically on social media. The diversity and size of the sample cannot and will not represent the broader population of Instagram users. Additionally, because the data was obtained through snowball sampling and not representative of the broader Instagram user base, it is possible that some results might not be generalized to all demographic groups.

Therefore, the findings of this research should be interpreted with caution when attempting to apply them to real life algorithmic experiences.

Furthermore, it is important to acknowledge the potential external factors that could have influenced participants' perceptions of the study. Because the topic of this study deals with socially sensitive issues such as race, colorism, and algorithmic bias, pre-existing opinions or experiences to colorism or trust in social media feeds could seriously influence how they respond to the conditions regardless of the experimental manipulation. Future studies similar to this should recognize and understand the cognitive and emotional context participants bring to the study. Social desirability also plays a role here. Participants may have given more socially acceptable answers when questioned about their perceptions of platform trust and fairness in hopes of appearing more unbiased. This is important to acknowledge in future research, especially in research involving race and discrimination. Another important external factor that could affect ecological validity is the familiarity and usage of the platform. Perceptions of Instagram's algorithm can vary depending on how much a participant uses the platform. For instance, someone that rarely uses Instagram may not easily recognize the biases in its algorithm. Future studies should include familiarity and usage of a platform when performing a similar study. Lastly, the sample of this thesis consisted of participants from multiple nationalities and ethnic backgrounds, thus participants might interpret skin tone diversity differently than others. To narrow this down, future research could focus on one demographic characteristic such as age, gender, nationality, or education level.

The second limitation highlights the manipulation check. Although the manipulation check had a statistically significant association between the assigned conditions and participants noticing diversity in the feeds, there was still 40% of the participants in the lighter skin tone condition that still reported noticing diversity. This indicates that the manipulation was not perfectly controlled, and the internal validity could have been compromised. Some participants may have interpreted visual content in unintended ways or might not have attended to the intended cues of the mock feeds. Future research should utilize stimuli that is stronger or more clearly differentiates the conditions.

The third limitation concerns the measurement of friend group diversity. The scale used in this study focused more on attitudes towards racial differences and discomfort rather than actual social connectedness or the quality of cross-group friendships. This may not have fully captured the intended construct and might have affected the accuracy of the moderation analysis due to not reflecting participants' true engagement with racially diverse individuals in their social circles. To adjust this, the current study could have asked participants about their ethnicity or membership of marginalized groups. Future research could incorporate this and/or benefit from a measurement that not just captures attitudes or biases but directly measures the quality, closeness, and frequency of interpersonal relationships with racially diverse friends.

5.5 Conclusion

This thesis set out to explore the impact of colorist algorithmic bias on perceptions of platform trust and fairness on Instagram by answering the following research question: *"To what extent do colorist algorithmic biases impact users' perceived trust of and platform fairness of curated social media feeds?"*. The results demonstrate that racial representation influences these perceptions. While there was no decline in trust and fairness following the exposure to predominantly lighter-skinned feeds, there was significant improvement after exposure to the feeds with diverse skin tones. Interestingly, the findings also suggest that the racial diversity in one's interpersonal relationships does not impact these outcomes, suggesting that identifying algorithmic bias can still occur despite their cross-group friendships.

By incorporating colorism into this study, these findings provide a comprehensive understanding of how social inequalities can be reinforced through personalized algorithmic systems. Decisions made by algorithms (neutral or not) are shaped by the training data and design choices by their developers. Thus, it is important to understand that real-life inequalities can be embedded in digital environments and can greatly influence user experiences in subtle and powerful ways. As digital media continues to shape how we see ourselves and others, this research contributes to ongoing discourse about digital media ethics and algorithmic fairness. Ensuring transparency and inclusivity in algorithmic designs (in all algorithmic systems, regardless of social media) is not only a technical challenge but also a moral responsibility. In conclusion, colorist algorithmic bias is not an abstract concern; it has tangible effects on how platforms are perceived, how users engage with them, and how fairness is understood in digital spaces. Future research should continue to question how visual and cultural representation interact with data-driven technologies to shape public perceptions, trust, and equity in the digital age.

References

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., & Gao, G. (2022). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, *12*(1), 100702. https://doi.org/10.1016/j.hlpt.2022.100702

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., & Shen, K. (2021). Algorithmic bias in data-driven innovation in the age of AI. *International Journal of Information Management*, *60*, 102387. https://doi.org/10.1016/j.ijinfomgt.2021.102387

Babbie, E. R. (2017). *The basics of social research* (7th ed.). Cengage Learning.

Baha, K. (2023). Algorithmic Bias: Causes and effects on marginalized communities. *University of San* Diego, 1–18. https://doi.org/10.22371/04.2023.001

Bahns, A. J., Springer, L. S., & Carla. (2015). Fostering diverse friendships: The role of beliefs about the value of diversity. *Group Processes & Intergroup Relations*, *18*(4), 475–488. https://doi.org/10.1177/1368430214566893

Bijou, C., & Colen, C. G. (2022). Shades of health: Skin color, ethnicity, and mental health among Black Americans. *Social Science & Medicine*, *313*, 115387. https://doi.org/10.1016/j.socscimed.2022.115387

Bodle, R. (2019). Predictive algorithms and personalization services on social network sites. In Routledge eBooks (eds), *Writing in a Technological World* (1st ed., pp. 273–288). Routledge. https://doi.org/10.4324/9780429507014-19

Butkowski, C., Humphreys, L., & Mall, U. (2022). Computing colorism: skin tone in online retail imagery. *Visual Communication*, *23*(2), 289–308. https://doi.org/10.1177/14703572221077444

Cabiddu, F., Moi, L., Patriotta, G., & Allen, D. G. (2022). Why do users trust algorithms? A review and conceptualization of initial trust and trust over time. *European Management Journal*, *40*(5), 685–706. https://doi.org/10.1016/j.emj.2022.06.001

Calvano, E., Haghtalab, N., Vitercik, E., & Zhao, E. (2024). Algorithmic content selection and the impact of user disengagement. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2410.13108

Chattopadhyay, S. (2019*). Fair-Unfair: Prevalence of colorism in Indian matrimonial ads and married women's perceptions of skin-tone bias in India*. Bowling Green State University. https://www.proquest.com/docview/2316417366?%20Theses&fromopenview=true&pq-origsite=gscholar&sourcetype=Dissertations%20

Chen, C., & Sundar, S. S. (2024). Communicating and combating algorithmic bias: effects of data diversity, labeler diversity, performance bias, and user feedback on AI trust. *Human-Computer Interaction*, 1–37. https://doi.org/10.1080/07370024.2024.2392494

Childs, K. M. (2022). "The Shade of It All": How Black women use Instagram and YouTube to contest colorism in the beauty industry. *Social Media + Society*, *8*(2). https://doi.org/10.1177/20563051221107634

Chuan, C., Sun, R., Tian, S., & Tsai, W. S. (2024). EXplainable Artificial Intelligence (XAI) for facilitating recognition of algorithmic bias: An experiment from imposed users' perspectives. *Telematics and Informatics*, *91*, 102135. https://doi.org/10.1016/j.tele.2024.102135

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). SAGE Publications.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th edition). SAGE Publications.

Davies, K., Tropp, L. R., Aron, A., Pettigrew, T. F., & Wright, S. C. (2011). Cross-group friendships and intergroup attitudes. *Personality and Social Psychology Review*, *15*(4), 332–351. https://doi.org/10.1177/1088868311411103

Dixon, J., Tropp, L. R., Durrheim, K., & Tredoux, C. (2010). "Let them eat harmony." *Current Directions in Psychological Science*, *19*(2), 76–80. https://doi.org/10.1177/0963721410363366

Edelman, B., Luca, M., & Svirsky, D. (2017). Racial discrimination in the sharing Economy: evidence from a field experiment. *American Economic Journal Applied Economics*, *9*(2), 1–22. https://doi.org/10.1257/app.20160213

Eg, R., Tønnesen, Ö. D., & Tennfjord, M. K. (2022). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, *9*, 100253. https://doi.org/10.1016/j.chbr.2022.100253

Eslami, M., Vaccaro, K., Lee, M. K., On, A. E. B., Gilbert, E., & Karahalios, K. (2019). User attitudes towards algorithmic opacity and transparency in online reviewing platforms. *AMC Digital Library*, 1–14. https://doi.org/10.1145/3290605.3300724

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8). https://doi.org/10.1111/phc3.12760

Fehr, B., & Harasymchuk, C. (2022). Conceptions and the experience of friendship in underrepresented groups. *Personal Relationships*, *29*(3), 451–487. https://doi.org/10.1111/pere.12431

Fu, Y., Wang, S., King, C., & Chu, Y. (2018). The influence of an international festival on visitors' attitudes toward diverse cultures. *Academica Turistica*, *12*(2), 127–141. https://doi.org/10.26493/2335-4194.11.127-141

Hampton, L. M. (2021). Black feminist musings on algorithmic oppression. *arXiv (Cornell University)*. https://doi.org/10.1145/3442188.3445929

Head, A. J., Fister, B., & MacMillan, M. (2020). Information literacy in the age of algorithms. *Student experiences with news and information, and the need for change*. Project Information Literacy Research Institute.

Hern, A. (2021, August 10). Student proves Twitter algorithm 'bias' toward lighter, slimmer, younger faces. *The Guardian*. https://www.theguardian.com/technology/2021/aug/10/twitters-image-cropping-algorithm-prefers-younger-slimmer-faces-with-lighter-skin-analysis

Hunter, M. (2007). The persistent problem of colorism: Skin tone, status, and inequality. *Sociology Compass*, *1*(1), 237–254. https://doi.org/10.1111/j.1751-9020.2007.00006.x

Hunter, M. (2012). The consequences of colorism. In Hall, R. (eds), *The Melanin Millennium* (pp. 247–256). Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4608-4_16

Jacoba, C. M. P., Celi, L. A., Lorch, A. C., Fickweiler, W., Sobrin, L., Gichoya, J. W., Aiello, L. P., & Silva, P. S. (2023). Bias and Non-Diversity of Big Data in Artificial Intelligence: Focus on retinal Diseases. *Seminars in Ophthalmology*, *38*(5), 433–441. https://doi.org/10.1080/08820538.2023.2168486

Jarvis, C., & Quinlan, M. (2022). IVF So white, so medical: Digital normativity and algorithm bias in infertility on Instagram. *Human-Machine Communication*, *5*, 133–149. https://doi.org/10.30658/hmc.5.6

Keyes, L., Small, E., & Nikolova, S. (2020). The complex relationship between colorism and poor health outcomes with African Americans: A systematic review. *Analyses of Social Issues and Public Policy*, *20*(1), 676–697. https://doi.org/10.1111/asap.12223

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085x.2021.1927212

Love, T. (2022). #TheDarkestShade: Exploring skin tone labor and the role of colorism in the cultural messaging around Black beauty influencers. *Deep Blue (University of Michigan)*. https://doi.org/10.7302/6012

Lu, J., Chua, S. N., Kavanaugh, J. R., Prashar, J., Ndip-Agbor, E., Santoso, M., Jackson, D. A., Chakraborty, P., Raffoul, A., & Austin, S. B. (2024). Impact of digital advertising policy on harmful product promotion: Natural language processing analysis of skin-lightening ads. *American Journal of Preventive Medicine*. *67*(6), 811-819. https://doi.org/10.1016/j.amepre.2024.08.006

Lury, C., & Day, S. (2019). Algorithmic personalization as a mode of individuation. *Theory Culture & Society*, *36*(2), 17–37. https://doi.org/10.1177/0263276418818888

Monk, E. P. (2021). The unceasing significance of colorism: skin tone stratification in the United States. *Daedalus*, *150*(2), 76–90. https://doi.org/10.1162/daed_a_01847

Mourali, M., Novakowski, D., Pogacar, R., & Brigden, N. (2025). Public perception of accuracy-fairness trade-offs in algorithmic decisions in the United States. *PLoS ONE*, *20*(3), e0319861. https://doi.org/10.1371/journal.pone.0319861

Neuman, W. L. (2014). Experimental research. In *Social research methods: Qualitative and quantitative approaches* (7th ed., pp. 281-313). Pearson.

Noble, S. U. (2013). Google Search: Hyper-visibility as a means of rendering Black women and girls invisible. *InVisible Culture*. https://doi.org/10.47761/494a02f6.50883fff

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Page-Gould, E., Harris, K., MacInnis, C. C., Danyluck, C. M., & Miller, I. D. (2022). The intergroup perspective on cross-group friendship. In *Advances in experimental social psychology* (pp. 1–56). Elsevier. https://doi.org/10.1016/bs.aesp.2021.10.001

Paudyal, P., & Wong, B. W. (2018). Algorithmic opacity: Making algorithmic processes transparent through abstraction hierarchy. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting, *62*(1), 192–196. https://doi.org/10.1177/1541931218621046

Pessach, D., & Shmueli, E. (2023). Algorithmic fairness. In Rokach, L., Maimon, O., Shmueli, E. (eds). *Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook* (pp. 867-886). Springer, Cham. https://doi.org/10.1007/978-3-031-24628-9_37

Riccio, P., Colin, J., Ogolla, S., & Oliver, N. (2024). Mirror, mirror on the wall, who is the whitest of all? Racial biases in social media beauty filters. *Social Media + Society*, *10*(2). https://doi.org/10.1177/20563051241239295

Ryan-Mosley, T. (2021, August 15). An Ancient Form of Prejudice About Skin Color is Flourishing in the Modern Internet Age. MIT Technology Review. Retrieved December 2021, from https://www.technologyreview. com/2021/08/15/1031804/digital-beauty-filters-photoshop-photoediting-colorism-racism/?utm_source=Twitter&utm_medium=tr_social& tm_ campaign=site_visitor.unpaid.engagement

Shin, D. (2020). User perceptions of algorithmic decisions in the Personalized AI System: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media*, *64*(4), 541–565. https://doi.org/10.1080/08838151.2020.1843357

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Stockstill, C., & Carson, G. (2022). Are lighter-skinned Tanisha and Jamal worth more pay? White people's gendered colorism toward Black job applicants with racialized names. *Ethnic and Racial Studies*, *45*(5), 896–917. https://doi.org/10.1080/01419870.2021.1900584

TEDx Talks. (2024, February 28). *Error 404: Human Face Not Found / Robin Aisha Pocornie / TEDxAmsterdam* [Video]. Youtube. https://youtu.be/pVfvYYUkIcY?si=j8U2dkYL8NfBTnzX

Trammel, J. M. (2023). Artificial intelligence for social evil: Exploring how AI and beauty filters perpetuate colorism—Lessons learned from a colorism giant, Brazil. In Langmia, K. (eds). *Black Communication in the Age of Disinformation* (pp. 51–71). Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-27696-5_3

Vaske, J. J., Beaman, J., & Sponarski, C. C. (2016). Rethinking internal consistency in Cronbach's Alpha. *Leisure Sciences*, *39*(2), 163–173. https://doi.org/10.1080/01490400.2015.1127189

Wang, X., Zhang, Y., & Zhu, R. (2022). A brief review on algorithmic fairness. *Management System Engineering*, *1*(1). https://doi.org/10.1007/s44176-022-00006-z

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. *AMC Digital Library*, *656*, 1–14. https://doi.org/10.1145/3173574.3174230

Wu, W., Huang, Y., & Qian, L. (2023). Social trust and algorithmic equity: The societal perspectives of users' intention to interact with algorithm recommendation systems. *Decision Support Systems*, *178*, 114115. https://doi.org/10.1016/j.dss.2023.114115

Yang, H., Li, D., & Hu, P. (2024). Decoding algorithm fatigue: The role of algorithmic literacy, information cocoons, and algorithmic opacity. *Technology in Society*, 102749. https://doi.org/10.1016/j.techsoc.2024.102749

Yapo, A., & Weiss, J. (2018). Ethical implications of bias in machine Learning. *Proceedings of the . . . Annual Hawaii International Conference on System Sciences/Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2018.668

Appendix A

Master thesis

Start of Block: Default Question Block

Q1 Dear participant,

I would like to formally thank you for taking the time to participate in this survey. This research project is part of a Master's thesis that aims to **examine perceptions of social media feeds.**

You will be asked a few questions about your perception of personalized algorithms. Afterward, you will be presented with a few social media feeds. Please view them **carefully** and **take your time**. Please answer the questions truthfully, as there are no right or wrong answers. The survey will take approximately **5 minutes** to fill in.

Your responses will strictly be used for research purposes and will not be shared with any third parties, ensuring confidentiality. Additionally, your responses will be completely anonymous and will not be traceable. You are free to withdraw from this survey at any point with no consequences. If you have any concerns or further questions about this study, please feel free to contact **558813zk@eur.nl**

Please continue if you are at least 18 years old and consent to participate in this study.

○ I consent (1)

End of Block: Default Question Block

Start of Block: Social media use

1 Which social media platform(s) do you use at least once a week?

- ☐ Instagram (1)
- ☐ TikTok (2)
- ☐ Twitter/X (3)
- ☐ Facebook (4)
- ☐ Pinterest (5)
- ☐ Snapchat (6)
- ☐ Other (please specify) (7)

____

2 What type of content do you consume on social media? Please select the one(s) that apply

to you the most.

- ☐ Beauty (1)
- ☐ Fashion (2)
- ☐ Fitness (3)
- ☐ Lifestyle/vlogs (4)
- ☐ Cooking/Food (5)
- ☐ Storytimes/Commentary (6)
- ☐ Live streams (7)
- ☐ Podcasts (8)
- ☐ Gaming (9)
- ☐ Sports (10)
- ☐ Other (please specify) (11) ____

End of Block: Social media use

Start of Block:

Q28 You will now be presented with a series of statements about your perceptions of the algorithms used by Instagram.

End of Block:

Start of Block: Pretest: Trust

1 Please indicate the extent to which you agree with the following statements

I trust the recommendations by Instagram's algorithm.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

---

2 Recommended content through Instagram's algorithm is trustworthy.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

3 I believe that Instagram's algorithmic feed results are reliable.

- ○ Strongly disagree (1)
- ○ Disagree (2)
- ○ Somewhat disagree (3)
- ○ Neither agree nor disagree (4)
- ○ Somewhat agree (5)
- ○ Agree (6)
- ○ Strongly agree (7)

End of Block: Pretest: Trust

Start of Block: Pretest: Fairness

1 Please indicate the extent to which you agree with the following statements.

Instagram's algorithm show no favoritism and do not discriminate against people.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

---

2 I believe Instagram should be transparent about the data it collects for its recommendations, document that data properly, and regularly evaluate its fairness and accuracy.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

3 I believe Instagram's algorithm is made to be impartial and without prejudice.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

End of Block: Pretest: Fairness

Start of Block: Block 12

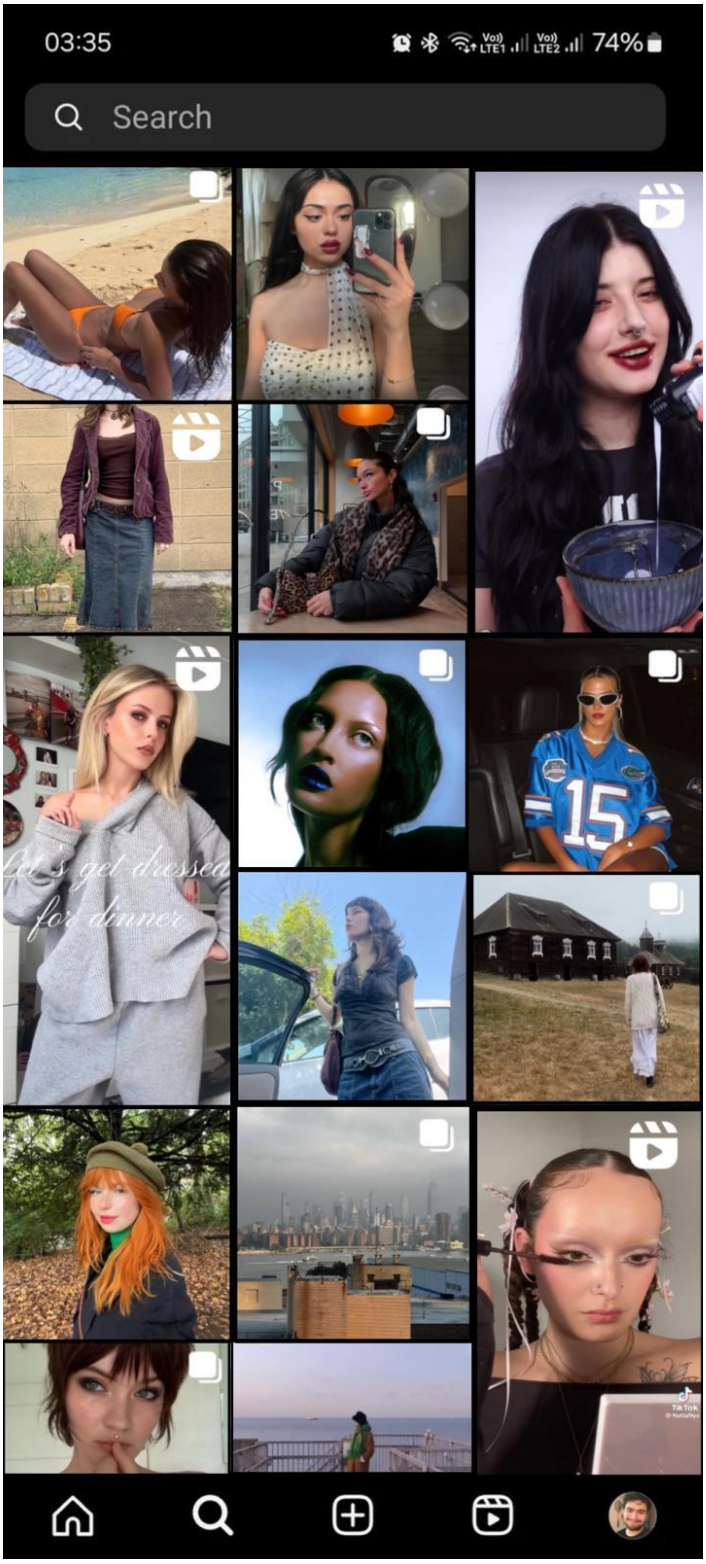
Q37 Next, you will be presented with **four** personalized feeds from Instagram. After viewing each feed, you will be presented with statements about what you saw.

End of Block: Block 12

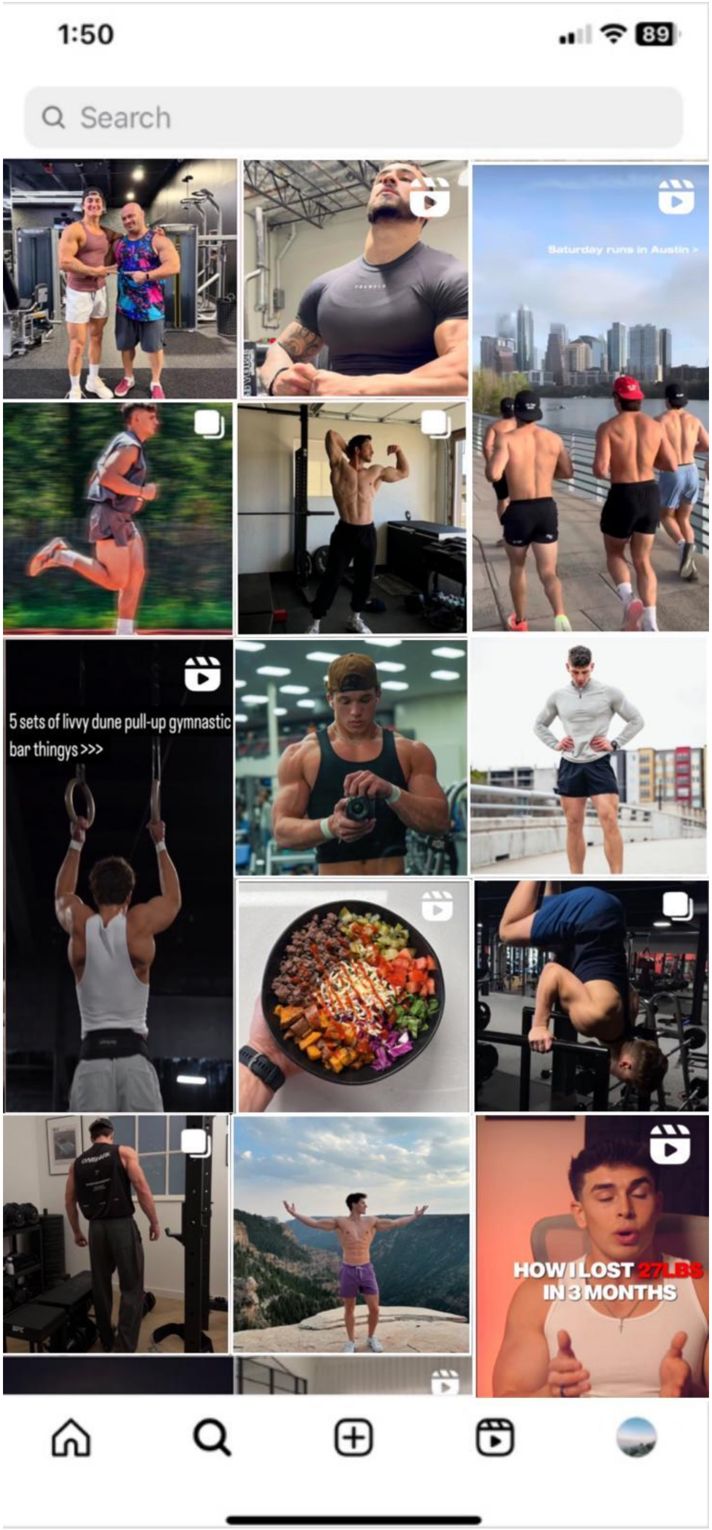
Start of Block: Mockup feeds: Lighter skin tones

Condition 1

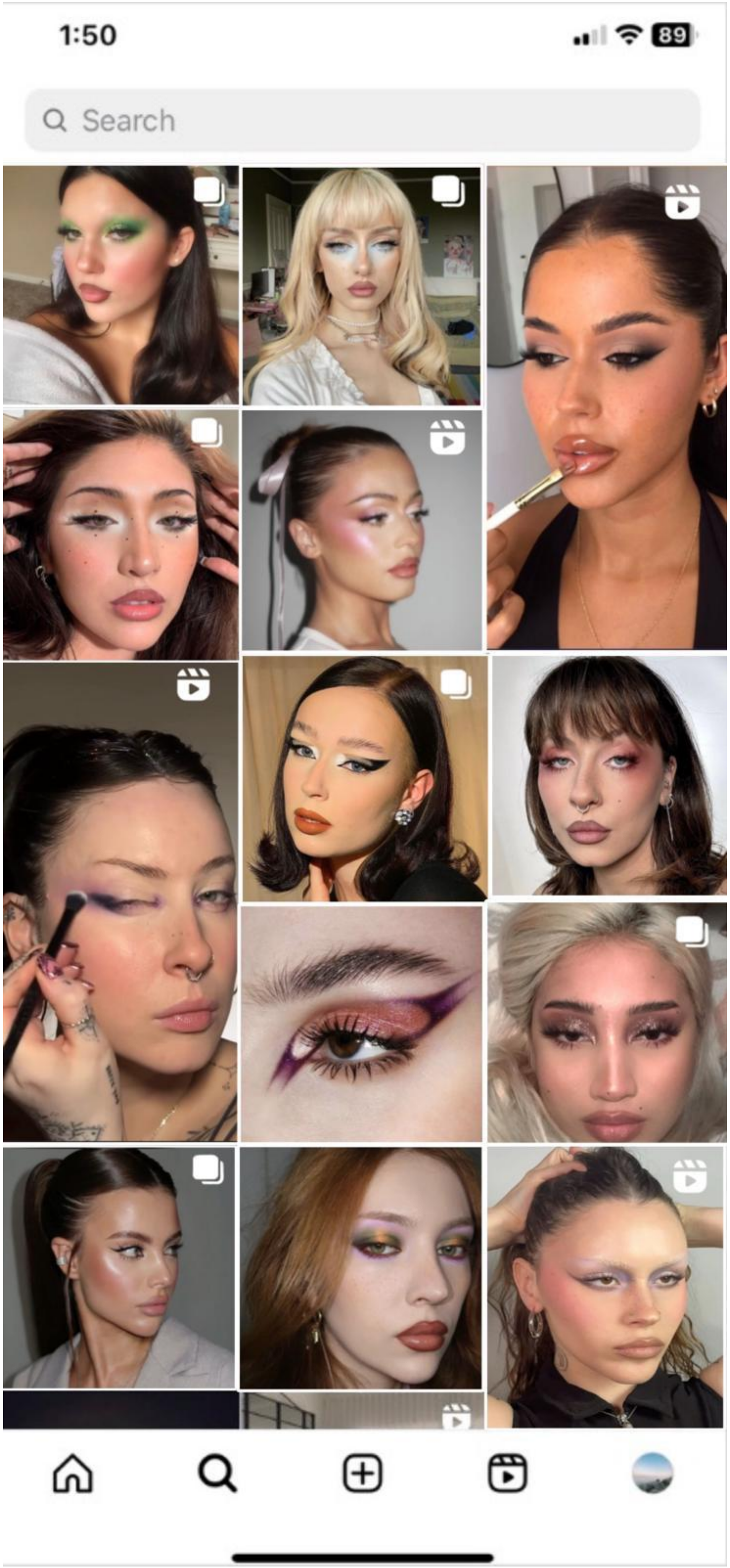
LighterSkin: Feed 1 Please take a look at the following feeds.



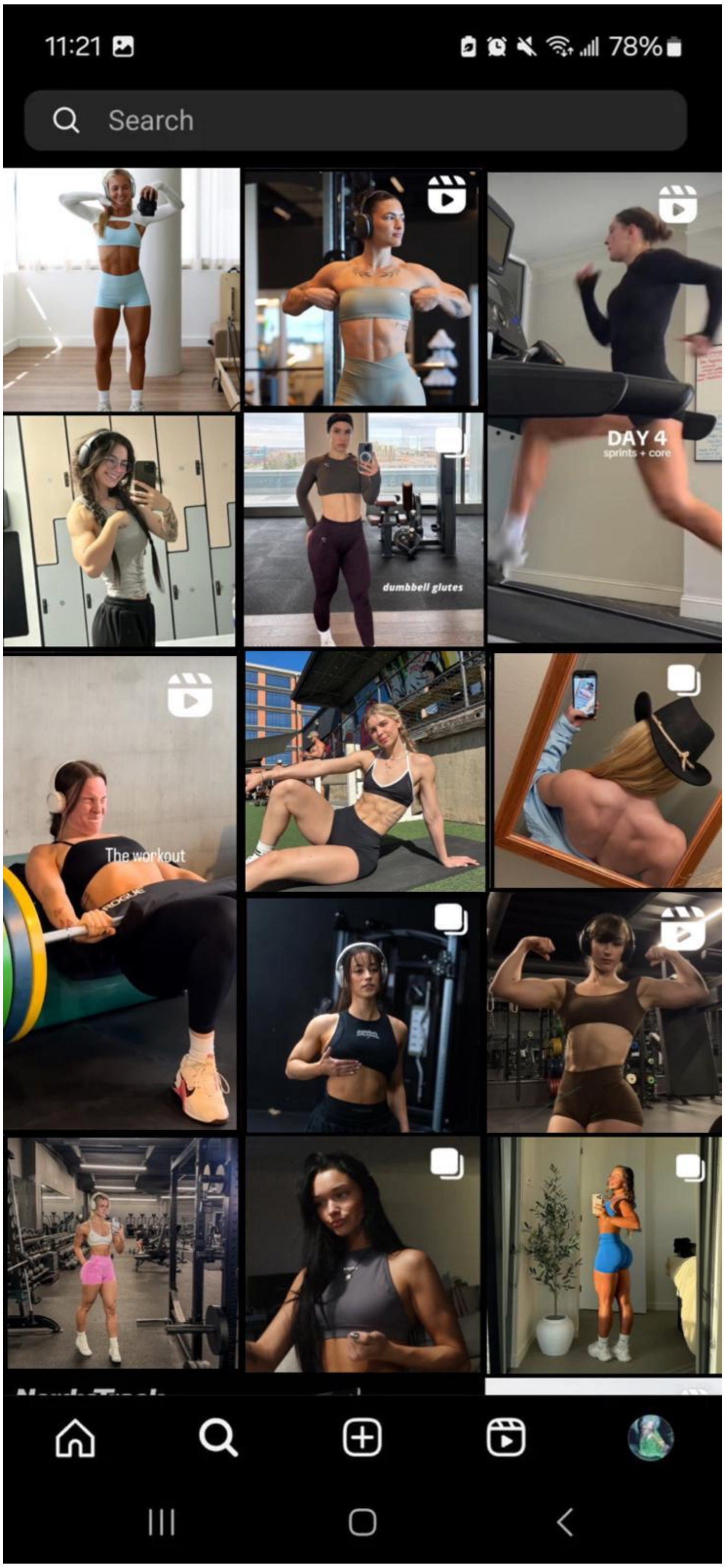
LighterSkin: Feed 2



LighterSkin: Feed 3



LighterSkin: Feed 4

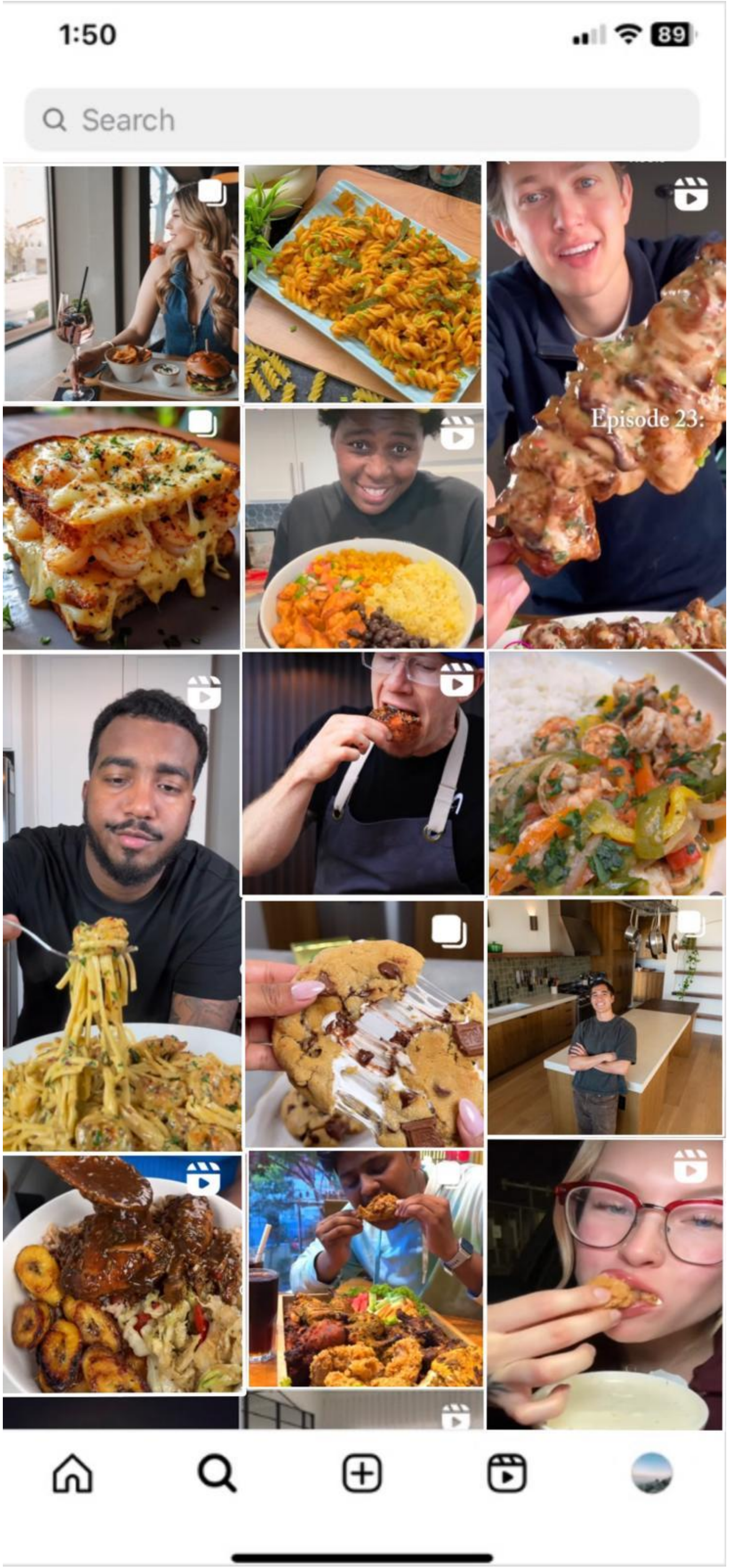


End of Block: Mockup feeds: Lighter skin tones

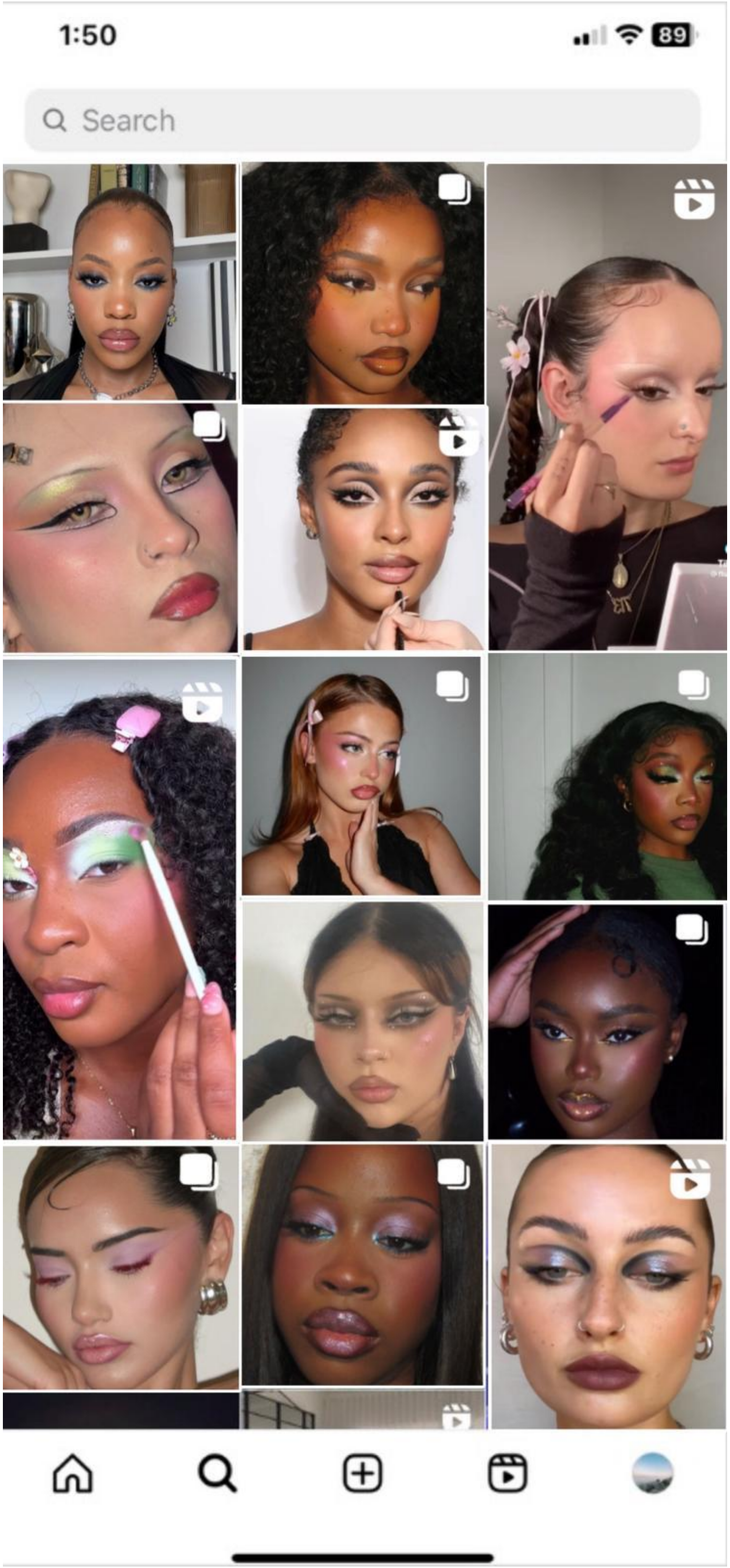
Start of Block: Mockup feeds: Diverse skin tones

Condition 2

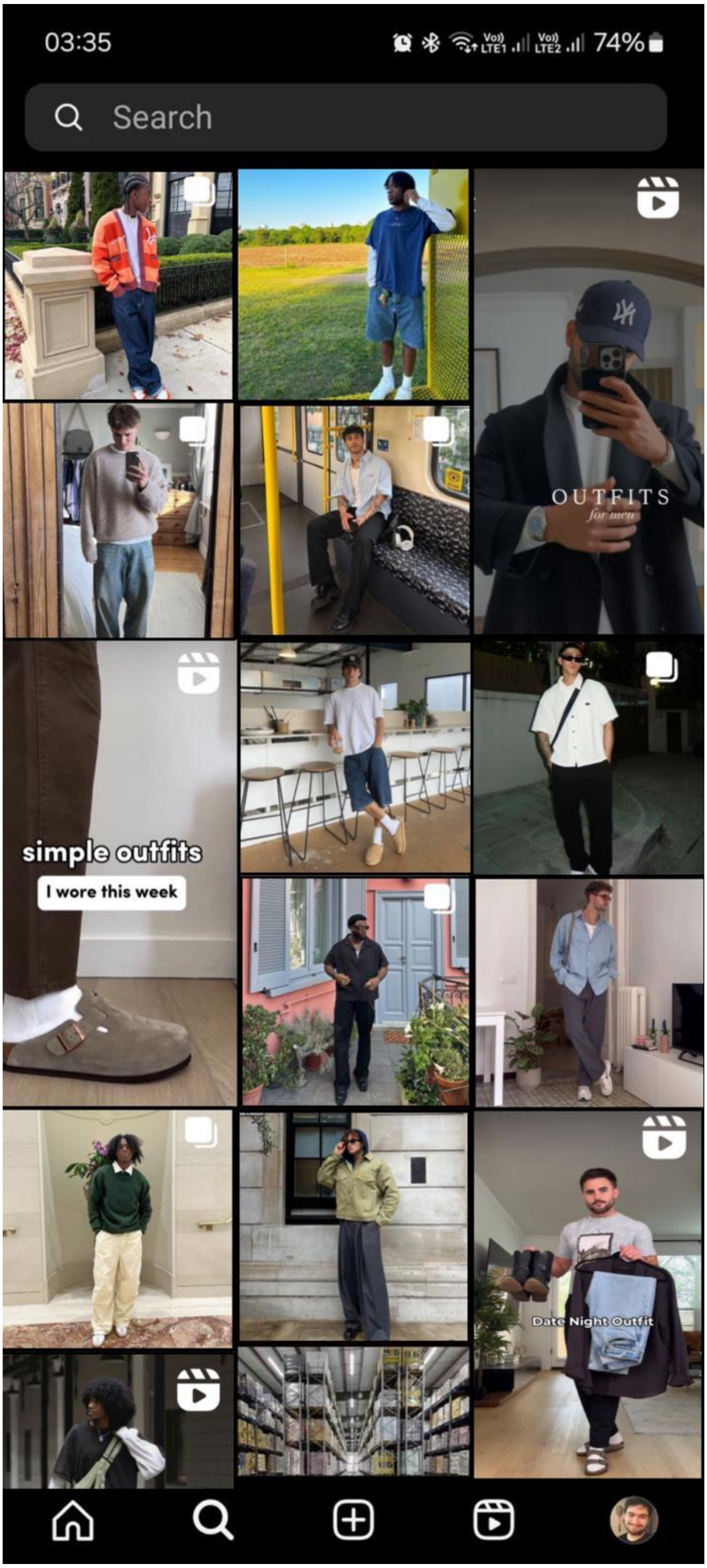
DiverseSkin: Feed 1 Please take a look at the following feeds.



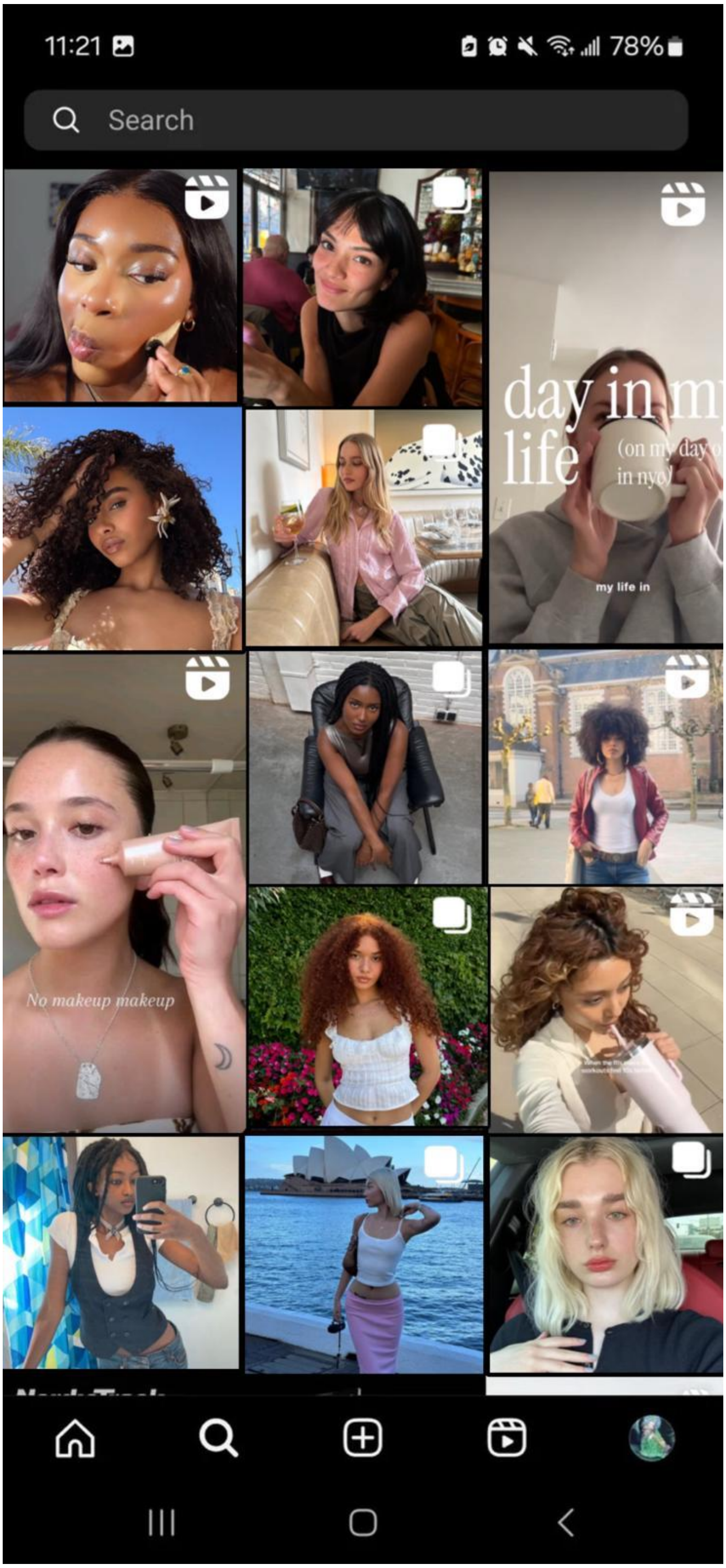
DiverseSkin: Feed 2



DiverseSkin: Feed 3



DiverseSkin: Feed 4



End of Block: Mockup feeds: Diverse skin tones

Start of Block: Post-test: Trust

1 Please indicate the extent to which you agree with the following statements

I trust the recommendations by Instagram's algorithm.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

2 Recommended content through Instagram's algorithm is trustworthy.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

---

3 I believe that Instagram's algorithmic feed results are reliable.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

End of Block: Post-test: Trust

Start of Block: Post-test: Fairness

1 Please indicate the extent to which you agree with the following statements

Instagram's algorithm show no favoritism and do not discriminate against people.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

---

2 I believe Instagram should be transparent about the data it collects for its recommendations, document that data properly, and regularly evaluate its fairness and

accuracy.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

---

3 I believe Instagram's algorithm is made to be impartial and without prejudice.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

End of Block: Post-test: Fairness

Start of Block: Manipulation check

1 Did you notice any diverse skin tones in the feeds?

○ Yes (1)

○ No (2)

End of Block: Manipulation check

Start of Block: Block 11

Q36 Great, you're almost finished! Now you will be presented with a few statements about your friend group(s).

End of Block: Block 11

Start of Block: Cross-group friendships

1 How many people in your social circle have a different ethnic background than you?

○ 0 (1)

○ 1 (2)

○ 2 (3)

○ 3 (4)

○ 4 (5)

○ 5+ (6)

2 Please indicate the extent to which you agree with the following statements.

Getting to know someone of another race is generally an uncomfortable experience for me.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

---

3 I am only at ease with people of my own race.

- ◯ Strongly disagree (1)
- ◯ Disagree (2)
- ◯ Somewhat disagree (3)
- ◯ Neither agree nor disagree (4)
- ◯ Somewhat agree (5)
- ◯ Agree (6)
- ◯ Strongly agree (7)

---

4 It's really hard for me to feel close to a person from another race.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

---

5 It is very important that a friend agrees with me on most issues.

- Strongly disagree (1)
- Disagree (2)
- Somewhat disagree (3)
- Neither agree nor disagree (4)
- Somewhat agree (5)
- Agree (6)
- Strongly agree (7)

---

6 I often feel irritated by persons of a different race.

◯ Strongly disagree (1)

◯ Disagree (2)

◯ Somewhat disagree (3)

◯ Neither agree nor disagree (4)

◯ Somewhat agree (5)

◯ Agree (6)

◯ Strongly agree (7)

End of Block: Cross-group friendships

Start of Block: Demographics

Age What is your age in years? (e.g., 22)

____

____

____

____

____

X→

country What country best represents your nationality?

▼ Afghanistan (1) ... Zimbabwe (1357)

Gender What is your gender?

- ◯ Male (1)
- ◯ Female (2)
- ◯ Non-binary / third gender (3)
- ◯ Prefer not to say (4)

Education What is the highest level of school you have completed or the highest degree you have received?

- ◯ Less than high school degree (1)
- ◯ High school graduate (2)
- ◯ Some college but no degree (3)
- ◯ Bachelor's degree (4)
- ◯ Master's degree (5)
- ◯ Doctoral degree (6)
- ◯ Professional degree (JD, MD) (7)

End of Block: Demographics

Appendix B

Table B1

Distribution scores of demographics in the lighter skin tones condition

	N	Trust (pre-test)	Trust (post-test)	Fairness (pre-test)	Fairness (post-test)
Age					
19	4	4.33	4.50	2.88	3.63
20	6	4.56	4.39	3.33	3.33
21	7	4.67	4.76	3.00	2.93
22	17	3.71	3.43	2.32	2.18
23	11	3.82	3.73	2.32	2.73
24	6	4.11	3.67	2.00	1.83
25	4	4.83	4.08	2.25	1.50
26	1	2.00	1.67	1.00	1.00
27	1	6.00	5.00	5.50	4.50
29	3	4.22	4.00	4.17	4.00
32	2	3.17	3.33	1.00	1.50
33	1	2.00	2.00	4.00	4.00
34	2	3.00	4.17	3.75	2.50
37	1	3.00	3.00	3.00	3.00
39	1	4.00	4.00	4.00	4.00
41	1	4.00	4.33	4.50	4.50
42	2	2.83	2.50	3.00	2.75
45	1	4.33	2.67	3.50	1.50
46	1	4.00	4.00	4.00	4.00
47	1	2.00	2.00	4.00	4.00
49	2	4.50	4.50	3.75	3.25
50	1	4.00	3.67	4.00	4.00
52	1	2.67	2.33	4.00	2.50
53	2	3.33	2.50	2.75	2.00
54	3	4.56	4.00	3.83	4.00
55	5	2.93	2.53	3.00	2.80
56	3	4.22	4.44	4.33	4.33
57	1	5.00	4.33	4.50	4.50

58	3	3.67	4.22	4.33	4.33
59	2	2.33	2.33	2.75	2.75
60	4	3.75	4.25	4.00	4.25
63	1	4.00	4.00	4.00	4.00
64	1	4.00	4.00	4.00	4.00
67	1	2.00	1.00	3.50	1.00
69	1	4.00	4.33	4.50	4.00
70	1	4.00	4.00	4.00	4.00
Nationality					
Bahamas	1	2.67	2.00	3.00	1.00
Belgium	2	2.00	2.33	1.50	1.25
Canada	1	2.00	1.67	2.00	1.00
Cape Verde	1	4.67	4.00	4.00	6.00
China	2	3.50	2.50	2.25	1.75
Estonia	2	4.00	4.17	3.50	3.25
Finland	1	3.67	3.67	1.00	1.00
France	4	3.92	3.08	2.13	1.38
Germany	2	5.17	5.50	2.50	1.75
Greece	2	5.33	5.00	2.25	2.75
India	3	5.89	5.33	3.00	3.17
Indonesia	2	4.83	4.83	2.25	2.75
Iraq	1	4.00	4.00	1.00	1.00
Ireland	1	5.00	2.00	1.00	1.00
Italy	3	4.22	4.00	3.50	3.33
Mexico	1	3.67	3.00	3.50	6.00
Netherlands	14	3.55	3.19	2.54	2.43
Pakistan	1	4.67	3.67	1.50	1.00
Poland	1	3.00	3.00	2.50	2.50
Portugal	1	2.33	3.67	2.00	3.00
Romania	1	3.67	3.33	2.00	2.00
Russia	1	2.00	3.00	3.00	3.00
Serbia	1	4.33	4.67	2.50	3.00
Singapore	1	5.00	3.00	1.50	2.00
Spain	1	5.33	6.00	6.00	6.50

Suriname	50	3.85	3.93	3.65	3.51
United Kingdom	2	3.67	2.33	1.50	1.00
Unites States of America	2	2.17	2.33	2.50	3.00
Gender					
Male	21	3.52	3.46	3.05	3.07
Female	78	3.98	3.80	3.08	2.88
Non-binary	5	3.27	3.07	2.30	2.50

Table B2

Distribution scores of demographics in the diverse skin tones condition

	N	**Trust (pre-test)**	**Trust (post-test)**	**Fairness (pre-test)**	**Fairness (post-test)**
Age					
19	1	5.00	5.33	3.50	4.00
20	8	4.50	4.79	2.65	2.50
21	11	4.39	4.84	3.55	3.45
22	24	4.04	4.31	2.46	3.08
23	13	3.49	3.74	3.27	3.31
24	6	3.56	3.44	2.92	2.75
25	7	4.24	4.52	2.50	2.86
26	1	4.67	5.00	4.50	4.00
28	3	3.67	3.78	2.83	3.33
29	1	4.33	4.67	3.00	3.50
31	1	4.00	4.00	3.50	3.50
33	1	2.33	2.00	1.00	1.00
40	1	7.00	6.00	6.00	6.00
43	1	4.00	4.67	1.00	3.00
44	1	5.00	5.00	3.50	3.00
45	2	1.83	2.67	2.50	2.50
46	1	4.67	5.00	4.50	5.50
47	2	3.83	3.83	4.25	4.25

49	3	4.11	3.89	4.83	5.00
50	4	3.67	4.08	3.50	4.00
51	1	2.33	3.00	3.00	3.00
52	2	3.50	4.00	3.75	4.00
53	2	4.00	3.50	3.50	3.50
54	1	4.00	6.00	4.00	4.00
55	2	2.83	2.83	2.75	2.00
56	1	1.33	1.33	1.00	1.00
57	2	4.83	4.67	4.75	5.00
58	2	3.17	3.00	3.75	2.50
59	1	5.00	4.00	5.00	3.50
60	1	4.00	4.00	4.00	4.00
61	2	3.67	3.00	4.00	4.75
68	1	5.33	6.00	6.00	6.00
Nationality					
Albania	1	2.00	2.00	2.00	2.00
Armenia	1	3.67	4.00	3.00	5.00
Belgium	4	4.42	4.25	2.75	2.50
Brazil	2	3.17	3.50	2.75	3.25
Bulgaria	2	4.17	4.50	2.25	2.25
Canada	3	3.89	3.78	1.83	2.83
Croatia	1	6.00	6.00	3.00	3.00
Czech Republic	2	5.50	6.00	4.75	6.00
Estonia	1	4.33	2.33	2.50	1.50
Finland	2	4.50	2.33	2.50	1.50
France	7	4.05	4.19	2.57	2.50
Germany	4	3.00	3.50	2.50	2.25
Hungary	2	3.50	3.67	3.75	4.00
India	1	5.00	5.00	3.00	2.00
Indonesia	2	4.67	4.33	4.75	4.00
Ireland	1	4.67	3.33	2.00	2.50
Italy	3	4.22	4.67	2.00	2.17
Japan	1	4.00	5.00	3.50	5.00

Kazakhstan	1	5.00	5.00	3.50	4.00
Malaysia	1	6.00	6.00	6.00	6.00
Netherlands	16	3.54	3.46	2.97	3.38
Poland	1	5.00	4.33	1.00	1.00
Portugal	1	3.67	3.00	3.50	3.00
Romania	2	3.17	4.00	3.25	3.50
Suriname	42	3.94	4.32	3.67	3.83
Thailand	1	4.67	5.00	3.00	3.50
Ukraine	1	4.00	4.00	3.50	3.50
United Kingdom	2	3.17	3.50	2.25	2.25
United States of America	3	4.89	4.67	3.67	3.33
Gender					
Male	17	3.90	4.04	3.15	3.21
Female	88	3.99	4.16	3.18	3.40
Non-binary	4	3.67	3.67	2.75	2.75

Appendix C

Declaration Page: Use of Generative AI Tools in Thesis

Student Information

Name: Zahra Kraag

Student ID: 558813

Course Name: Master Thesis CM5000

Supervisor Name: Prof. Dr. Marc Verboord

Date: 26/06/2025

Declaration:

Acknowledgment of Generative AI Tools

I acknowledge that I am aware of the existence and functionality of generative artificial intelligence (AI) tools, which are capable of producing content such as text, images, and other creative works autonomously.

GenAI use would include, but not limited to:

- Generated content (e.g., ChatGPT, Quillbot) limited strictly to content that is not assessed (e.g., thesis title).
- Writing improvements, including grammar and spelling corrections (e.g., Grammarly)
- Language translation (e.g., DeepL), without generative AI alterations/improvements.
- Research task assistance (e.g., finding survey scales, qualitative coding verification, debugging code)
- Using GenAI as a search engine tool to find academic articles or books (e.g.,

☒ I declare that I have used generative AI tools, specifically [Name of the AI Tool(s) or Framework(s) Used], in the process of creating parts or components of my thesis. The purpose of using these tools was to aid in generating content or assisting with specific aspects of thesis work.	☐ I declare that I have NOT used any generative AI tools and that the assignment concerned is my original work. Signature: [digital signature] Date of Signature: [Date of Submission]

Extent of AI Usage

☒ I confirm that while I utilized generative AI tools to aid in content creation, the majority of the intellectual effort, creative input, and decision-making involved in completing the thesis were undertaken by me. <u>I have enclosed the prompts/logging of the GenAI tool use in an appendix.</u>

Ethical and Academic Integrity

☒ I understand the ethical implications and academic integrity concerns related to the use of AI tools in coursework. I assure that the AI-generated content was used responsibly, and any content derived from these tools has been appropriately cited and attributed according to the guidelines provided by the instructor and the course. I have taken necessary steps to distinguish between my original work and the AI-generated contributions. Any direct quotations, paraphrased content, or other forms of AI-generated material have been properly referenced in accordance with academic conventions.

By signing this declaration, I affirm that this declaration is accurate and truthful. I take full responsibility for the integrity of my assignment and am prepared to discuss and explain the role of generative AI tools in my creative process if required by the instructor or the Examination Board. I further affirm that I have used generative AI tools in accordance with ethical standards and academic integrity expectations.

Signature: 

Date of Signature: 26/06/2025

Appendix D

Prompt: Use of AI to summarize a paragraph from an academic article

copy pasted paragraph “please summarize this text for me and emphasize the examples”

ChatGPT:

This passage discusses how training datasets used in AI applications can introduce algorithmic bias, leading to unfair outcomes. The key causes of bias include:

1. Sample Selection Bias – If the training dataset is not representative of the target population, the algorithm’s decisions can be skewed.
    - Example: Amazon’s AI-based recruitment tool was abandoned because it discriminated against female applicants due to the lack of female data in the training set.
    - Example: Apple’s credit card algorithm offered lower credit limits to women compared to men, highlighting gender bias in financial AI models.
2. Out-Group Homogeneity Bias – When data is incorrectly grouped, developers might assume that individuals within a category are more similar than they actually are.
3. Size of the Training Dataset – A small dataset increases bias since it lacks diversity and can reinforce existing inequalities.
4. Reinforcement Learning Fairness Issues – AI struggles with counterfactual data, meaning it cannot predict how someone would behave in a situation it has never seen.
    - Example: A loan application algorithm cannot predict if someone denied a loan would have repaid it, leading to unfair credit assessments.
5. Recommendation Engine Bias – AI-based content recommendations can create blind spots, limiting users’ ability to discover new content.
    - Example: Some products or services may never be recommended to users, making it harder for them to find certain items.
    - Example: Personalized filters may reinforce polarization in rating data, causing biased feedback loops.

Main takeaway:

Bias in AI stems from how training datasets are created and used, impacting everything from hiring and credit decisions to content recommendations. These biases can reinforce discrimination and limit fair access to opportunities.

Prompt: Use of AI to explain something to make it more comprehensible

copy pasted sentence form text “what is being said here?”

ChapGPT:

This sentence is explaining how content-based filtering works in recommendation systems.

Let me break it down simply:

👉 Content-based filtering recommends items to you based on the features or characteristics of items you've already liked or interacted with.

In other words:

- It doesn't rely on what other people like.
- It looks at what you personally liked before, and then finds similar items.
- These similarities are based on specific features — like genre, color, price, topic, etc.

Example:

If you watched a romantic comedy on Netflix and rated it highly, the algorithm might recommend other romantic comedies (same genre), or movies with the same actors, tone, or plot themes — because these features match the one you liked.

So the sentence is saying:

> Content-based filtering analyzes the properties of content you've liked and then finds new items with similar properties to recommend to you.