Re:framing
Migrants in the
European Media

Report and visualisation of media representation dynamics

Project title	Re:framing Migrants in the European Media		
Start date	01/02/2022	**Duration**	15 months
Project URL			
Contractual due date	31/08/2022	**Actual submission date**	29/08/2022
Nature	R= Document, report	**Dissemination level**	PU= public
Authors	Francesca Trevisan, Mitchel Njoki, Gemma Galdon Clavell		
Contributors	Evren Yalaz, Patricia Vazquez, Emilia Paesano		
Reviewers	Menno Weijs, Badria Zeino-Mahmalat		

This project has received funding from the *European Commission DG CNECT* under grant agreement No LC01727412.

EXECUTIVE SUMMARY

Social media is a virtual space where people exchange their opinion, communicate with friends, look for information and read news. Contrary to the traditional media, such as newspapers and television, social media allows people to create their own online identities, establish their own networks, tell their stories, and contribute to the public debate, bringing it to new directions. With social media being increasingly part of people's routine, its use and dynamics have become controversial in terms of violation of fundamental rights, racism, sexism and lack of transparency and accountability. The platforms' models of governance, and the new forms of content creation and sharing can lead to the severe intrusions on fundamental rights such as the right to equality and non-discrimination, the right to privacy and can escalate to new forms of systemic violence. The social media infrastructure is defined by opaque algorithms and obscure practices that reinforce stereotypes and create new forms of discrimination that are particularly harmful to the social groups that are marginalized by the power structure. Migrants and refugees are part of the community affected by these new forms of discrimination. Understanding and managing the particular risks connected to the use by, and the representation of refugees and migrants in social media platforms need to be a priority of regulations addressing the digital space. For an effective management and mitigation of these risks, it is fundamental to identify and understand how social media's socio-technical dynamics work, how they manipulate the representation of reality and affect the experience and representation of marginalised communities. This report contributes to this objective, and it is divided into four main sections.

Section 1 introduces the main topics that will be discussed in this report, it outlines the purpose of the "Re:framing Migrants in the European Media" project and explains how this report is aligned with it. It presents the approach on which the report is based and the key definitions of the lexicon used.

Section 2 identifies and discusses the key social media dynamics that affect the representation of marginalised populations on social media. It provides an overview of how social media algorithmic dynamics works, how they silence marginalised communities and where bias affecting them can arise. It illustrates stories of discrimination, silencing of voices and manipulation of reality. To this end, this section is structured around four themes: content moderation, shadowban, content selection and sharing, and targeted advertisement.

Section 3 focuses on the migrant and refugees' population. It shows how social media plays an important role in shaping new forms of migration and how this can represent a danger for migrants

and refugees. It outlines how migrants and refugees are represented on social media platforms such as Instagram, Pinterest, Youtube, Facebook and Twitter and how their representation is related to issues of misinformation, hate speech and incitement to violence. **Section 4** provides a conclusion, along with a summary of the content.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

The visibility of migrants and refugees in the European public sphere is almost always peripheral: they are often denied agency, voice, and the right to be represented as complex human beings. In the media, migrants' and refugees' stories are narrated by others with sensationalizing tones, political ends, stereotypical images, and stigmatizing frames (e.g. Eberl et al., 2018). CNECT is a 15-months project funded by the European Commission DG CNECT which aims to provide a comprehensive and holistic analysis of the causes of both the lack of and the misrepresentation of migrants and refugees in European media. The project is undertaken by a consortium of 6 partners from 4 European countries.

The core vision of the project is to change the current media narratives through assurance of appropriate representation of the migrant and refugee communities across Europe in an inclusive and empowering manner, providing for a space of fair and un-discriminatory self-representation. Social media are part of the media ecosystem that shapes migrants' representation and the way people discuss migration. Throughout the years, representation issues and dynamics in traditional media (e.g., newspapers, radio, television) have been amply discussed, while attention dedicated to the role of social media as spaces of representation is more recent. Social media socio-technical infrastructure plays a crucial role in processes of identity building, opinion making, and it creates new forms of expression, narration, and information. This becomes particularly problematic when we consider that social media platforms are developed and administered with little transparency and accountability. They rely on algorithms that are not public and on values that follow the market logic in their design, governance, and management. This means that social media needs to be profitable, efficient, and competitive. The saying "Move fast break things" describes well how social media platforms work: the product is on the market before safeguards are put in place. This has direct consequences for their users. In fact, while social media allowed better connectivity between people, its infrastructure defined also new forms of bias, misrepresentation, discrimination, and suppression of voices. This threatens marginalised communities that struggle to be fairly represented and to have their voice heard. In order to understand the lack of representation and the misrepresentation of migrants and refugees in the European scene, it is crucial to focus on the social and technical features that shape social media representation dynamics and harm marginalised groups. To achieve this purpose, this report presents how social media dynamics affect representation and voice of marginalised groups, it outlines the role of social media in migration and discusses how migrants are represented on social media.

1.1 Purpose and Scope

This report is entitled ***Report and visualization of media representation dynamics*** and explores the key media representation dynamics to identify the mechanisms behind social media that results in the current practices of discrimination and silencing of voices. It shows how social media is changing the migration phenomenon and how social media content represents migrants and refugees. The social media mechanisms that contribute to the misrepresentation, discrimination and silencing of marginalised groups presented in this report are essential to uncover structural inequalities that are entrenched in social media and reproduced by these platforms. This is the key to subvert the system, understand how to regulate it and create tools and practices that allow people to be protected while having their voice heard on these platforms.

The outcome of this report aligns with and contributes to the overall objective of the project in changing the current social media narratives through the assurance of appropriate media representation of migrant and refugee communities across Europe in an inclusive and empowering manner, providing a space where migrants can represent themselves.

1.2 Structure of the report

This report is divided into four main sections.

The current section -Section 1- is a general introduction to the report.

Section 2 reports media representation dynamics and uncovers the mechanisms behind social media that make current discrimination and silencing of voices possible. It shows how content moderation, algorithms that sort content in social media and targeted advertising affect marginalized groups.

Section 3 introduces the role of social media in migration. It collects the formal and informal literature on how social media influences migration processes and reviews the state of the art that investigated how social media represents migrants and refugees.

Section 4 presents a general conclusion.

1.3 LEXICON

Artificial Intelligence: according to the European Commission's High Level Expert Group on Artificial Intelligence, 2018 "Artificial intelligence (AI) refers to systems designed by humans that, given a complex goal, act in the physical or digital world by perceiving their environment, interpreting the collected structured or unstructured data, reasoning on the knowledge derived from this data and deciding the best action(s) to take (according to predefined parameters) to achieve the given goal. AI systems can also be designed to learn to adapt their behaviour by analysing how the environment is affected by their previous actions. As a scientific discipline, AI includes several approaches and techniques, such as machine learning (of which deep learning and reinforcement learning are specific examples), machine reasoning (which includes planning, scheduling, knowledge representation and reasoning, search, and optimization), and robotics (which includes control, perception, sensors and actuators, as well as the integration of all other techniques into cyber-physical systems)."

Algorithm - a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

Algorithmic Bias - algorithmic bias describes systematic errors in a computer system that create unfair outcomes, such as privileging one group over others.

Machine learning - Machine learning is a branch of Artificial Intelligence and Computer Science which uses data and algorithms to imitate the way that humans learn (IBM, 2020).

Migrant - There is no consensus on the exact definition of a migrant. The UN Migration Agency defines a migrant as "any person who is moving or has moved across an international border or within a State away from his/her habitual place of residence, regardless of (1) the person's legal status; (2) whether the movement is voluntary or involuntary; (3) what the causes for the movement are; or (4) what the length of the stay is." (UN, 2020).

Online toxicity - is an umbrella term for online hateful behaviour that manifests itself in forms such as online hate speech (Salminen et al., 2020).

Refugee - Refugees are people who have fled war, violence, conflict, or persecution and have crossed an international border to find safety in another country (UN, 2020).

Social Media - Social networking websites and applications that enable users to create and share content.

Socio-Technical System - a system that involves both humans and technology.

2. SOCIAL MEDIA REPRESENTATION DYNAMICS

Social media is composed of websites and applications that allow users to create content, share it and participate in social networking. Worldwide, people spend on average two and a half hours per day on social media (Biselli & Beckmann, 2020; Meaker, 2018) and , among the platforms available on the market, Facebook, YouTube, TikTok and Twitter are some of the most popular (Statista, 2022b). Social media is increasingly embedded in everyday communication and is an important source of news and information. Generally, people use social media to keep in touch with friends and family, establish new connections, find and create content, see what is being talked about, read news, engage in online discussion and purchase new items (GWI, 2022). Media outlets and journalists across Europe are also increasingly relying on social networking platforms to report news, promote stories and expand their audiences (Brems et al., 2017; Bruno, 2011).

> Over the years, social media has revolutionized the way information is produced, disseminated and consumed.

Social media platforms have the control over the information users are exposed to using opaque algorithms, based on business models, and informed by thousands of data about their users and the available content. For this reason, social media dynamics and the use of algorithms demands a closer inspection of how they work and what values they prioritize. What users see in social media is the result of the selection of all the content available for them to see after that the content that has been judged as “unacceptable” has been removed or demoted. Furthermore, the content that is shown is tailored to the users’ interests: personalization algorithms observe users' digital habits, predict their choices and decide what information to show users. Based on users' profiles (e.g., demographic features) and their activities (e.g. interactions with other users, posts), social media platforms also offer targeted advertisement. These dynamics follow the attention economy, which aim at keeping users' engaged. Consequently, social media users are exposed to a representation of the reality that is filtered and that tends to be in line with the users' interests and opinions.

> Researchers around the world have shown how online social networks and big data algorithms act as echo chambers where like-minded people are trapped in ideological circles that can harm democracy and representation (Barisone & Michailidou, 2017). This contributes to the spread of disinformation.

The social media dynamics described above shape new forms of discrimination. They can reinforce stereotypes and have direct implications on how social groups are represented. This issue is amplified by the insufficient platform oversight, transparency and regulatory frameworks that render social media platforms scarcely accountable. This section of the report aims at identifying the mechanisms behind social media that make stereotyping, current discrimination, and silencing of voices possible and will present the key issues related to the representation of vulnerable groups on social media. It will show the reader how social media dynamics reproduce structures of power that discriminate marginalised groups and how discrimination is embedded in computer codes and in artificial intelligence technologies. Section 2.1 provides an overview on algorithmic bias and how it affects social media platforms. Section 2.2 and 2.3 introduce two different forms of platform governance: content moderation and shadowban. Section 2.4 presents how content is selected and shared while section 2.5 looks at targeted advertisement. For each topic, the representation issues and sources of bias that further marginalise vulnerable communities will be outlined.

2.1 ALGORITHMIC BIAS

Algorithms, machine-learning (ML) and artificial intelligence (AI) are models that, depending on human's objectives, can make predictions, recommendations or decisions influencing both the real and virtual environment. AI and its outcomes are often perceived and presented as neutral, objective, and accurate but often they are not. They can create unfair outcomes, such as disadvantaging some groups or skewing people representation of reality towards stereotypical and discriminatory frames. There are different sources of bias that might affect the outcome of AI models.

> Algorithms and ML models are created by people who hold beliefs, values and assumptions that can be injected in the design of the algorithm and can also affect how the data used to train the algorithm is selected, collected, and coded.

In other words, algorithmic bias can derive both from the structure of the algorithm and the data used. An example of algorithmic bias is the one found in the recruiting algorithm used by Amazon (Dastin, 2018). In 2015, Amazon realized that its automated system to screen candidates CVs was misbehaving. The Amazon recruiting system scored job candidates from one to five stars and did not rate candidates in a gender-neutral way. This occurred because the model was trained with CVs submitted to the company over a 10-year period and the majority came from men. As a consequence, Amazon's automated system taught itself from the data that male candidates were preferred and penalised women's CVs. Another example is the bias found by the Gender Shades project (Buolamwini & Gebru, 2018) which uncovered that automated facial analysis algorithms from leading companies held a gender and skin type bias when performing gender classification.

The systems were showing discrepancies in error rates when classifying lighter women, darker women, lighter men and darker men, with dark skinned women reporting the highest error rates compared to lighter skinned men, who had more accurate results. In other words, the training data of the facial classification systems were overwhelmingly white, causing racial and gender bias in the classification results.

The data used by the algorithms and the human decisions that are behind their creation, can reflect historic inequalities, stereotypes and assumptions that compromise some social groups and manipulate reality. Furthermore, the social groups who are penalized by algorithmic systems are often those who are historically disadvantaged. Therefore, algorithms can drive polarization and exacerbate existing inequalities that are rooted in our social systems. By using and repeating past patterns and human stereotypes and assumptions, inequality can be automated and perpetuated.

> Unfair social structures are transferred to algorithms and ML models through the data they use and the structure of the algorithm itself.

Social media uses algorithms and AI in multiple ways. Algorithms determine how to show the content, whether it violates the platform's terms and should be removed or demoted[1], predict interests and personalize the content according to what is most relevant to the user. However, these algorithms have been shown to be problematic in multiple ways. Social media algorithmic audits have shown issues (Bandy, 2021) in terms of:

- **discrimination:** where algorithms' output impact people unevenly depending on their gender, age, race, location, socioeconomic status, or intersectional identity.
- **distortion:** where algorithms present information in a way that distorts reality. For example, it might focus on some political views.
- **exploitation**: when algorithms inappropriately use users' data or sensitive information.
- **misjudgment**: when the algorithms make incorrect predictions or classifications and this might lead to discrimination, distortion of reality and exploitation.

For example, Twitter image cropping algorithm was found to be racist as it was focusing automatically on white faces over black faces (Hern, 2020). Twitter used this image cropping algorithm to automatically crop images to prevent them taking up too much space on the main feed and to allow multiple pictures to appear on the same page. The same algorithm was also found to discriminate against gender, age, weight, and non-western languages. It was found to favour Latin over Arabic text; to prefer thinner, younger looking people and women (e.g. male gaze) ; and to be biased against people with white hair (Knight, 2021). Another case of racism occurred in Zoom, where a black person

[1] Demotion on social media refers to the suppression of visibility of users' content.

using a virtual background on a call had his head removed as it was not recognized by the algorithm (Dickey, 2020). Personalisation and amplification algorithms can distort reality by unequally unevenly amplifying some voices and political views (Huszár et al., 2021) or by demoting content that is considered "undesirable". A good example of social media data exploitation is the Cambridge Analytica scandal, with Facebook inappropriately harvesting its users' data without the users' consent for political ends. Bias in social media can also be amplified in other ways. For example, in Zoom the most dominant speakers are also the ones who are most prominently featured in the Zoom gallery view. As participants join a zoom call, their thumbnail picture is added to the grid of participants in order of their arrival but then participants are reordered to the front as they speak (Rankin & MacDowell, 2021). This becomes problematic when considering that men are more likely to speak in class than women (J. J. Lee & Mccabe, 2021), and that black students have been reported to interact less in class (Eddy & Hogan, 2014). As a consequence, by showing the faces of those who speak the most, Zoom's algorithm reproduces the unequal participation that we would see in a class.

Despite companies tend to test their models to identify biases before they are used (Chowdhury, 2021), biases keep on being spotted. One of the challenges of identifying biases in social media algorithms is that the code they use is not public. In the name of privacy of their users and intellectual properties, social media companies do not share details of their algorithms. Consequently, models are hard to penetrate, which hampers research and adds an extra level of difficulty to knowing how these platforms work and how they take decisions.

> Keeping social media big tech companies accountable is hard as the sector is not properly regulated and largely not transparent.

Most of the times -such as the Twitter image cropping case-, complaints come from users who notice how platforms misbehave. In other cases, audits are carried out either independently by the research community (e.g. Cen & Shah, 2021), on civil societies requests (Murphy & Cacace, 2020) or internally (Huszár et al., 2021). This report approaches social media discrimination step by step. Sections 2.2 and 2.3 will outline how two different forms of platform governance -content moderation and shadowban- work to promote or demote users' content and will show how they can cause discrimination and manipulation of reality that affect the representation of marginalised communities.

2.2 CONTENT MODERATION

Every day, an unimaginable quantity of content is generated on online platforms like Facebook, Twitter, Instagram or TikTok. Online platforms rely on Community Standards or Terms of Services to regulate the behavior of their end users and moderate the content published. These Terms of Services are designed to protect human rights, prevent misinformation, harm (e.g. hate speech, violence, extremism, copyright infringements) and generally counter online toxicity.

> Terms of Services are privately developed, vary across platforms and are generally stricter than national laws as social media companies want to protect themselves from legal liabilities.

So far, the regulation and control of content published on online platforms has been left to the responsibility of the platforms. However, in July 2022 the EU Parliament and Council adopted a new regulation of online platform -the Digital Services Act[2]- to guarantee a safe digital space where fundamental human rights are protected. The Digital services act aims to establish stronger safeguards to ensure information is processed in a non-discriminatory way, in respect of fundamental rights and for a more transparent "notice and action" procedure where users are empowered to report harmful content online and challenge platforms' decisions.

Content moderation is the process that screens, monitors, and filters user-generated content against the platform specific guidelines to determine whether the content can be published or not.

> Once a piece of content has been moderated, it could be removed, banned, downgraded, or demonetized.

The goal of content moderation is to reduce the harmful content, safeguard users and maintain the reputation and credibility of the platform. Moderation is crucial to maintain a safe online environment and can be applied to different kinds of content: text, images, video, and live streaming. However, the

[2] https://digital-strategy.ec.europa.eu/it/node/27

content moderation operations enforced by online platforms are not transparent (MacCarthy, 2021): it is not known, nor compulsory this far to know- how content is detected, removed or demoted. Furthermore, companies have private documents for moderators that operationalise community guidelines with great details. These content moderation guidelines are not public, vary by country (Ryan et al., 2020) and change over time in relationship to shifting policies and norms within the company (e.g. Jaso, 2022). This section aims to give an overview of the main content moderation procedures to highlight different features of content moderation that can endanger the representation of marginalised communities.

Content moderation can be **automatic -** where ML algorithms screen, flag and remove content or **manual** when the screening, flagging and removal of content is performed by humans. However most of the times, content moderation procedures involve a combination of both automatic and human moderation. Content moderation can also happen at different points in the time of the publication process. The content can be screened against the community guidelines before it goes live on the platform (**pre-moderation**) or after its publication (**post-moderation**). Content moderation can also be **reactive** or **proactive**. **Reactive moderation** is dependent upon the users to report the content. For example, social media websites have the report option which allows users to report any content that they feel is inappropriate or that does not comply with the community standards. On the other hand, **proactive moderation** involves the use of machine learning filters that flag the content that might be automatically removed or go into a queue for review by human moderators.

Every time that content that seems to violate the Community standard is identified - either by machine learning filters or by being flagged by a user-, the content might be automatically removed or checked by a human being against the standard and removed if it does not abide by the community rules. In Facebook, for example, posts can be flagged by users or machine learning filters. Some of them are automatically dealt with from flag to removal, while some others go to a queue to be analysed by human moderators that often need to take controversial choices. While in the past posts were reviewed chronologically, machine learning techniques are used to order the queue and prioritize flagged content based on: its virality, severity and likelihood to break the rule (Vincent, 2020). Moderating online content implies a series of operational as well as ethical challenges. From an operational perspective, the enormous quantity of content that is produced real time worldwide, requires a great operational capacity to be flagged, analyzed and removed efficiently and effectively. From an ethical perspective, across platforms there is no consensus on the content that needs to be regulated or on who should be in charge of deciding the criteria for content moderation. As the

sections below will show, these challenges negatively impact fundamental rights like freedom of speech and expression, and harm the representation of some social groups.

2.2.1 Algorithmic content moderation

AI in content moderation is seen as the solution to the operational pressure caused by the increasing quantity of online content to moderate Algorithmic content moderation involves various techniques from statistics and computer science that aim to identify, match, predict or classify some piece of content.

> Algorithmic systems used for moderation can either match newly updated content against an existing database -is the image having the same features of another image?- or classify content that has no correspondent in the database -Is this content hate speech?.

Systems to **match** content generally use "hashing" which refers to the process of transforming a piece of content into a string of data meant to identify that unique piece of content. **Classification** tools for content moderation involve machine learning. These tools are trained with annotated datasets (e.g. images, text) depicting banned behaviors in order to identify them on the platform and remove them. They are able to automatically filter images that are prohibited by the community standards, block users violating guidelines and create blacklists that contain words, phrases and keywords to enable quicker detection and removal of content (Walker, 2021).

For the moderation of **images**, automated content moderation uses image processing algorithms trained to identify areas within images and create categories (e.g. harmful content) based on some chosen criteria. **Natural language processing (NLP)** algorithms are a classification tool that is used to moderate content. Natural language processing algorithms process and analyze large amounts of language data to predict the meaning of the text, the tone and whether it belongs to specific categories (e.g. hate speech) (Duarte et al., 2017). Text classification is used to assign categories to the text based on its sentiment or content. To classify the content, NLP tools employed by social media are trained with data labelled by human coders and then machine learning models use these annotated data to learn patterns about the content. For example, a model that detects hate speech learns from the training data the words that occur frequently in examples labeled as "hate speech" so

that they will be able to detect and potentially remove hate speech content. If the content is not automatically removed, it is passed to human moderators and the moderated content can be re-used as training data for the content moderator model.

However, information is often conveyed through different modalities: a post can have a picture, some texts and also some comments from other users. Some platforms (e.g. Facebook) have created models that understand content across modalities and violation types (Schroepfer, 2019).

2.2.2 Representation issues with content moderation

The COVID-19 pandemic, with the disruption of content moderation workforce and increased spread of misinformation, accelerated social media reliance on **automated tools**, resulting in an increase of **content moderation mistakes** as online content was erroneously flagged and removed (Heilweil, 2020).

> The inability of automatic tools to understand the context and the nuances of languages are factors that increase the probability that digital platforms delete content of marginalised groups.

The removal of content produced by marginalised populations contributes to generate exclusion, censorship, auto-censorship and social apathy in the digital environment. This dynamic reinforces and reproduces existing power structures that leave behind marginalised communities, making them invisible. For example, tools for automated text analysis have a limited ability to parse the meaning of human communication or to detect the motivation of the speaker. Natural language processing struggles to understand creative use of language (sarcasm, irony, humour, metaphors, creative social media spells of words, emoticons, abbreviations). Decisions based on automated social media content analysis risk to over-censor and further marginalize vulnerable groups that are already discriminated (Duarte et al., 2017). Furthermore, NLP models are trained on data that can be biased and the bias can be reflected in their outcome if it is not corrected (Davidson et al., 2019; Sap et al., 2019). This bias could lead to the moderation of content that misinterprets and censors the speech of certain groups such as marginalized groups or those with minority views. In this regard, many NLP tools work effectively only on text written in English. Therefore, the use of NLP tools on non-English content can compromise the outcomes for non-English speakers. NLP tools often struggle with variations in dialect and language, and this can result in less accuracy for minority populations. Demographic factors such as gender, ethnicity, race and location are associated with different language patterns. The

enforcement of Terms of Service by automated content moderators can disproportionally censor people of color, women, and other marginalized groups. In two studies published in 2019, researchers showed that AI intended to identify hate speech can also amplify bias. Sap et al., (2019) found that tweets written in African American English are up to two times more likely to be labelled as offensive compared to others. Davidson et al., (2019) using five different sets of Twitter data (155,800 tweets) showed similar evidence of racial bias as the system tended to predict at a higher rate that tweets in African American English were abusive.

Content moderation is often referred to as the gatekeeper of information: it is the moderation procedure that determines what can and cannot be seen.

> Community standards and moderation guidelines that companies keep secret are designed from point of views that show no sympathy for non-normative views, diverse social contexts, and realities.

It becomes particularly problematic when **profit priorities, politics and ideology affect content moderation rules and guidelines**. As a consequence, moderation decisions have direct implications for equality, free expression and to the free circulation of information.

In this regard, platforms have been increasingly accused of censorship and suppression of content relating to political issues and they have also been accused of not going far enough in capturing hateful and harmful content. Meta content moderator decisions have been under scrutiny different times in the past. Leaked documents revealed that the company was actively ignoring warnings from the integrity teams and did not do enough to mitigate social harm. For example, internal research findings indicating that Instagram was worsening body images issues for one in three teenage girls and that it was negatively impacting on their mental health were kept secret (Gayle, 2021). International and national politics also has been shown to affect Meta content moderation. Recently It was reported that since the beginning of the conflict between Russia and Ukraine, Meta made several content policy revisions (Jaso, 2022). Meta applied some temporary changes to its hate speech policy for users in Russia, Ukraine and Eastern-European countries, allowing them to post calls for death to Vladimir Putin and Alexander Lukashenko (Reuters, 2022). In a statement, Meta said “As a result of the Russian invasion of Ukraine we have temporarily made allowances for forms of political expression that would normally violate our rules, like violent speech such as ‘death to the Russian invaders’. We still won’t

allow credible calls for violence against Russian civilians," (Reuters, 2022). Meta standards for content were put under the spotlight also in 2020, when Mark Zuckerberg championed the commitment to free speech as a reason not to act on incendiary posts from Donald Trump - which were deemed by Facebook employees as attempts to incite violence - (Byers & Abbruzzese, 2020) implying that violence could be used against looters. However, at the same time, Facebook deactivated the account of dozens of Tunisian, Syrian and Palestinian activists and journalists that were documenting human rights abuses in the region (Solon, 2020). There are also cases in which social media platforms deliberately take down the post and account, often in coordination with the government. Instagram has been found to censor and ban Iranian media, individual journalists, human rights advocates and activists related to the killing of general Soleimani reportedly to comply with the U.S sanctions laws (Zakrezewski, 2020). For example, Instagram banned a post of an Iranian journalist who was a critic of the government but wrote that Soleimani's killing was "contrary to the principles of international law".

National activists' profiles have also been hit by content moderations. Black Lives Matter accounts have also been silenced and prevented from going live (Silverman, 2020) and users' posts talking about racism have been often censored as hate speech (Guynn, 2019). Furthermore in 2016, Facebook was accused to change history for having removed the iconic 1972 photo picturing children escaping the Nalpam bombs during the Vietnam war for child nudity. The "Nalpam Girl" picture, which won the Pulitzer Prize and has a high historical value, shows some children and soldiers escaping from a Napam bomb explosion on the background. The picture was removed because it contained a 9 years old naked girl and child nudity was against the platform community standards. After a community protest due to the historical value of the picture, Facebook decided to allow the image back on the platform. Facebook and Instagram moderation of nudity has been condemned by activists, after they banned pictures of female vs. male nipples or picture of women breastfeeding and after mastectomy (T. Gillespie, 2018).

Byte Dance (TikTok) was also accused of suppressing Black creators (Kelly, 2020). TikTok creators posting Black Lives Matter content reported to have it taken down, muted or hidden from viewers (McCluskey, 2020). TikTok was removing videos by Black activists talking about police brutality as they were "violating community guidelines" (Joslin, 2020). To address users' complaints, ByteDance launched a "creator diversity council", reviewed the moderation practices and said it was standing in solidarity with the black community. However, even after these measures, black creators reported "noticeable declines in viewership and engagement on their videos after posting content in support of the Black Lives Matter movement or noticed recent instances where they felt that TikTok's community guidelines weren't being fairly applied to Black creators" (McCluskey, 2020). TikTok also

blocked the viral video made by Feroza Aziz who disguised her Xinjiang video as a make-up tutorial. In the video, the teenager was raising awareness on how the Chinese government was accused of sending Uighurs and other minorities to internment camps. The video was removed and the user was blocked for posting any more videos and TikTok explained it as a human moderation error that should not have happened and promised to release a transparency report (Kuo, 2019). Another research showed that using the same queries, the proportions of bans for spelling, violation of guidelines and hateful behaviour in TikTok changed depending on the country where the VPN located the user indicating that moderation guidelines differ depending on where the content is produced (Tracking.Exposed project, 2022) a result that was also confirmed by ASPI researchers (Ryan et al., 2020).

2.3 SHADOWBAN

Shadowban is a term coined by social media users to refer to another form of platform governance that is much less researched because it is harder to detect and often denied by social media platforms. A shadowban happens when a user or the content created is silenced or blocked without any notification to the user.

> With shadowbans, content and profiles are still active and visible to the owner, but invisible to the others. This form of censorship limits the reach of potentially "problematic content and causes likes, comments and the overall engagement with the profile to drop.

It can be unclear to users whether their post or profile are visible and, in case, why they are not. When content or profiles are shadowbanned, the user is left to reverse-engineer its social media algorithm to understand what happened and how to secure visibility (Are, 2021). Shadowban offers social media platforms an **obscure moderation tools** that downgrade content and accounts of users that are considered "borderline". Borderline contents or accounts are those that do not directly infringe the community standard. They do not meet the moderation criteria to be removed, but they do not cross the acceptability line set by the platforms. As such, the content gets demoted and silenced without the creator being aware of it.

Who gets shadowbanned? Accounts of artists, human rights activists, political commentators, and ordinary users have been shadowbanned by different platforms. Instagram shadowban targeted

women bodies (Are, 2021), activists for disability and for the LGBTQIA+ community for posting explicit content (Impronta, 2020). The Intercept found TikTok internal moderation documents which instructed moderators to ban ideologically undesirable content and to algorithmically punish users who were considered unattractive (Biddle et al., 2020). Reportedly, TikTok instructed to shadowban posts created by users deemed too ugly, poor or disabled from the platform and to censor political speech in TikTok live streams. TikTok moderators were told to suppress users with "abnormal bodies", "ugly facial looks", "too many wrinkles" or in "slums, rural fields" and "dilapidated housing". Furthermore, content that was endangering the "national honor and interests" were punished with a ban. In 2018, The Guardian reported that TikTok was suppressing the content mentioning Tiananmen Square, Tibetan independence and banned the religious group Falung Dalong (Hern, 2019). As reported by the Guardian, the guidelines that instructed the shadowban of this content was under rules that seemed to be of general purpose such as banning "high controversial topics such as separatism, religion sects conflicts, conflicts between ethnic groups, for instance exaggerating the Islamic sects conflicts, inciting the independence of Northern Ireland, Republic of Chechnya, Tibet and Taiwan and exaggerating the ethnic conflict between black and white". TikTok was also banning a specific list of 20 "foreign leaders or sensitive figures" including Kim Jong-il, Kim Il-sung, Mahatma Gandhi, Vladimir Putin, Donald Trump, Barack Obama, Kim Jong-un, Shinzo Abe, Park Geun-Hee, Joko Widodo and Narendra Modi. Another report from NetzPolitik showed that TikTok algorithmically limited the reach of users that were disabled, overweight or LGBTQIA+ as an effort to prevent bullying. For some users, their content was not shown outside their native content, for others the content was not included in the "For You" page. These restrictions were applied to users who were deemed "susceptible to bullying or harassment based on their physical or mental condition", according to documents obtained by NetzPolitik.org, including "facial disfigurement, autism, Down syndrome, [or] disabled people or people with some facial problems". Other users were added to a list based on their "high risk" of bullying and a great number of the people listed had rainbow flags in their biographies and described them as lesbian, gay or non-binary, who found that hashtag referring to homosexuality, to the LGBTQIA+ community and other political hashtag were shadowbanned in different countries. The same report pointed out how other users who posted political content on TikTok were shadowbanned (Ryan et al., 2020).

There are different ways to shadowban accounts or their contents. To demote content, social media might:

- assign a lower relevance score to reduce the visibility of content that does not meet the bar of removal under their policy and might be offensive (e.g. Meta, 2021);

- prevent shadowbanned accounts and content to appear when they are searched (e.g. Stack, 2018; Tracking.Exposed project, 2022);
- prevent shadowbanned accounts to be found in feeds that are not part of their social network;
- silence shadowbanned content when shared by users that are part of the shadowbanned profile network.

Are (2021), who carried out an autoethnography on the shadowban of her own Instagram account, conceptualised the shadowban cycle: due to social media inability to prevent online harms and the fear of being overly regulated by government, tech companies find easy targets (e.g. content considered problematic) to conflate in harmful categories (sex trafficking, child abuse).

2.4 Conclusion on platform governance and representation issues

Content moderation practices and shadowban define the suitability of content and decide on the value of the digital content. To protect themselves from legal liabilities, platforms tend to over-censor their users' content. Furthermore, with the increasing quantity of online content and the pressure of governments to regulate the online environment, social media platforms are increasingly turning to automated content moderation. Automated content moderation poses different challenges for the representation of marginalised communities. The data used to train automatic tools to classify the online content are often produced by the dominant groups and culture, which fail to represent marginalised voices and perspectives. This makes social media platforms less able to deal with content posted by minorities and marginalised communities. The inability of these tools to understand nuances of language and context might result in an erroneous removal or can erroneously leave offensive and harmful content up on social media. In addition to this, terms of services and moderation guidelines are often impacted by ideological factors that affect the representation of marginalised populations and non-normative and diverse views.

To sum up, when a content is flagged, removed or demoted there are different possible scenarios:

- the post was considered as not complying to the rules and community standards;
- the post was deemed inappropriate due to political lobbying or ideology;
- the post was erroneously removed for human error;
- the post was erroneously removed for algorithmic bias.

These scenarios hold important issues and raise **important ethical concerns** on the platforms' policing in determining users and content visibility and how these affect fundamental rights of their users. Social media are a tool of cultural production (Poell et al., 2022) and have the power to re-write social norms and define what is socially acceptable, what should be seen, shared, debated and what should be silenced instead.

> By doing so, content moderation exercises a form of disciplinary power that regulate the behaviors of individuals and priorities some versions of reality that aligns with platforms' business interests by keeping people engaged and disfavour marginalised populations, activists and human rights advocates which representation, autonomy and free speech rights is undermined.

This is particularly threatening when thinking that there are billions of users that everyday use social media platforms to tell their stories and engage with news. Figure 1 visually summarizes how content moderation dynamics work and the multiple sources of bias that can affect users and more harshly, marginalised communities.

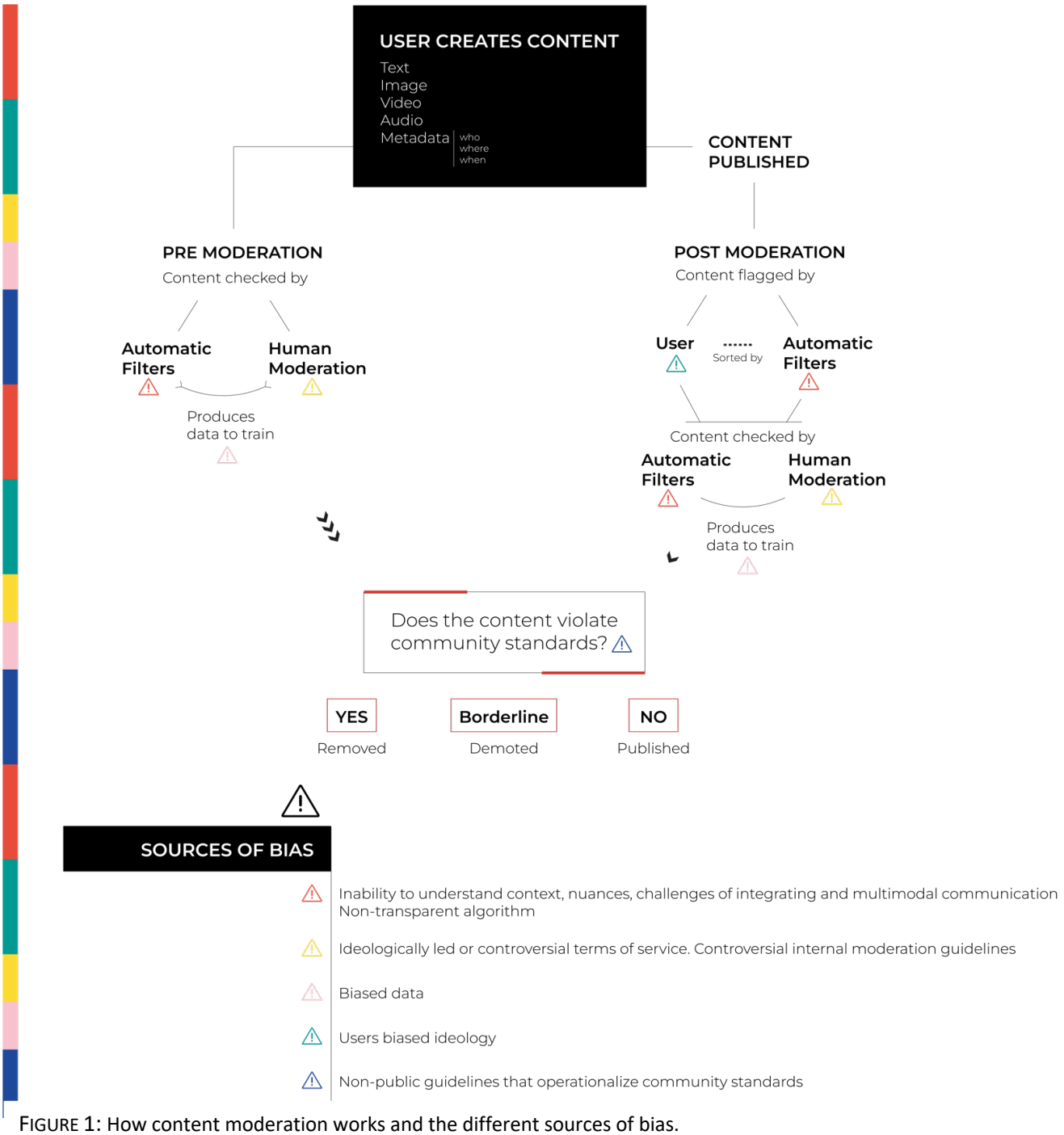

FIGURE 1: How content moderation works and the different sources of bias.

2.5 Content selection and sharing

After the content has been filtered through content moderation procedures, algorithms sort the content in users' timelines or news feed.

> Social media platforms are generally very vague about the algorithms they use for their news feed: big tech companies never reveal the core characteristics of the algorithms they use to sort users' content.

The goal of content selection algorithms is to select content for each user in order to maximize the engagement and attention -hence revenue- received; to do so, it must learn about the topics or groups the user is most interested in. Social media algorithms place users in environments characterised by selective exposure to information, that is filtered through one's own social network and reinforced by **recommendation** and **personalisation** algorithms. Personalisation and recommendation algorithms filter the content and consider users' characteristics and past behaviors to personalize the content to display. For example, they sort user's contents based on the accounts a user follows, on the comments posted, content shared, content created, and consumption time. Often, content is classified into different groups which are defined by attributes. For instance, news stories can have a political leaning (e.g., conservative or liberal), and a topic (e.g., politics, entertainment). The algorithm selects a piece of content to display to a given user, and receives feedback in the form of whether they engage with the content (e.g. click on it, purchase). To learn about the topics or groups the user is most interested in, the process is often modeled to update the probability distribution (from which one selects content), accordingly to feedback given. Positive feedback (click) will increase the probability of engaging with similar content, while negative feedback will diminish the probability of that user engaging with similar content in the future. As the content selection algorithm learns more about a user, the corresponding probability distribution begins to concentrate on a small subset of topics; this results in polarization, where the feed is primarily composed of a single type of content.

> These probabilistic models can affect users' sense of the self, their behaviour and make social media addictive.

These algorithms use different types of data:

- **Content data**: algorithms are able to identify and distinguish the features of the content published in order to make suitable recommendations to other users;
- **User data** which include users' demographic features (gender, age, career), its interests and behaviour in these platforms;
- **Scenario data** which refers to the record of user preference shifts based on different scenarios based, for example, on the user's geolocation.

For example, Facebook **news feed ranking algorithm** predicts what posts people might be most interested in and place them at the top of feeds. **News feed ranking** in Facebook is divided into four steps (Meta, 2021). First of all, Facebook algorithm computes an **inventory** that includes all the posts that a user could see: advertisements, users' posts and group activities. Then, the algorithm gathers information (**signals**) about the posts such as who posted it, how the users interacted with the person who created the post, whether it is a photo, video or text. These signals are then used to make **predictions** about the post based on how likely it is relevant to the users, whether the users might interact with it. Finally, the algorithm creates a **relevance score** for each post based on the signals and predictions. Posts with higher scores will be more likely to be of interest to the user so they will be placed at the top. Twitter timeline works very similarly and uses **ranking algorithms** to show personalised content to users. The tweets that users' see are "A stream of Tweets from accounts you have chosen to follow on Twitter, as well as recommendations of other content we think you might be interested in based on accounts you interact with frequently, Tweets you engage with, and more."(Twitter, 2022). On the other hand, TikTok algorithms appear to expose the content to an initial pool of viewers and then, based on video performance, it exposes the videos to other viewers considering users profile characteristics to do targeted amplification of the content. Also, Youtube recommendation system aims at helping viewers in finding videos they might want to find based on viewers watch history and videos engagement to maximise users' satisfaction[3].

2.5.1 Representation issues for content selection

These recommendation and personalisation algorithms are problematic for different reasons. Recommendation algorithms pose ethical challenges in terms of inappropriate content, users' data privacy, autonomy and personal identity, opacity, fairness and transformative power on society (Milano et al., 2020). They have been shown to be related to issues of amplification and silencing of

[3] https://www.youtube.com/watch?v=9Fn79qJa2Fc

various content (Huszár et al., 2021), polarisation (Conover et al., 2021; Weber et al., 2013), and create self-reinforcing biases and filter bubbles which can isolate people in their ideological views. Social media filter bubbles are also favoured by homophily: users with similar interests are more likely to be connected on social media, interact with each other and be exposed to similar information (Solomon et al., 2019). For example, Huszár et al., (2021) showed that ideological similarity between political parties increases their social media connectivity. Amplification and polarisation have been observed in several platforms. One of the most famous cases is Facebook amplification of hate speech that facilitated the genocide of Rohingya Muslims in Myanmar. Facebook was found to admit that the platform was used to incite violence and spread harmful and racially inflammatory content in Myanmar and did not do enough to prevent it (Meta, 2018a). After an independent human right assessment that confirmed the role that Facebook played in the escalation of religious and ethnic tension in the region, Facebook changed its policies to comply with human rights principles. They started to engage with the Myanmar local community to better understand the local context, employed native content moderators, improved their models to detect hate speech in Burmese (Meta, 2018b). Despite this, Global Witness pointed out that the ability to detect hate speech in Burmese is still very poor, as advertisements containing hate speech toward the Rohingya community were still accepted by Facebook for publication (Global Witness, 2022). Huszár et al., (2021) recently carried out the most comprehensive audit of Twitter recommender system by analysing 6.2 million news articles and millions of Twitter users. They found that Twitter did amplify some political content. The study in fact revealed that mainstream political right enjoyed more amplification than the political left and algorithmic amplification favoured right-wing news sources. We must point out that right-wing parties and news sources are also those who tend to discriminate marginalise communities. Hasell, (2021) further showed that partisan (vs. non partisan) news -which elicits more emotional response from the users- are shared more frequently in Twitter. In other words, partisan and emotional news content tend to be amplified in social media. This in turn, might affect how information is created in the online environment and increase the sensationalisation of content and can also polarise users. Mozilla (2021) audited YouTube's recommendation system through a crowdsourced activity that showed how the content among YouTube recommended videos was the one that was most signaled as regrettable by the users. Users found that YouTube was recommending them misinformation, inappropriate content, violent and racist content, hate speech, harmful content and conspiracy theories that could lead them spiraling into the darkest corners of the internet. Furthermore, the countries that were hit harder by regrettable recommended content were the ones who were not English speakers, suggesting that the recommendation algorithm works worse for those

who do not speak English as first language. Tik Tok's recommendation algorithm has been shown to spiral users down to videos promoting eating disorders and discriminatory content (Dias et al., 2021).

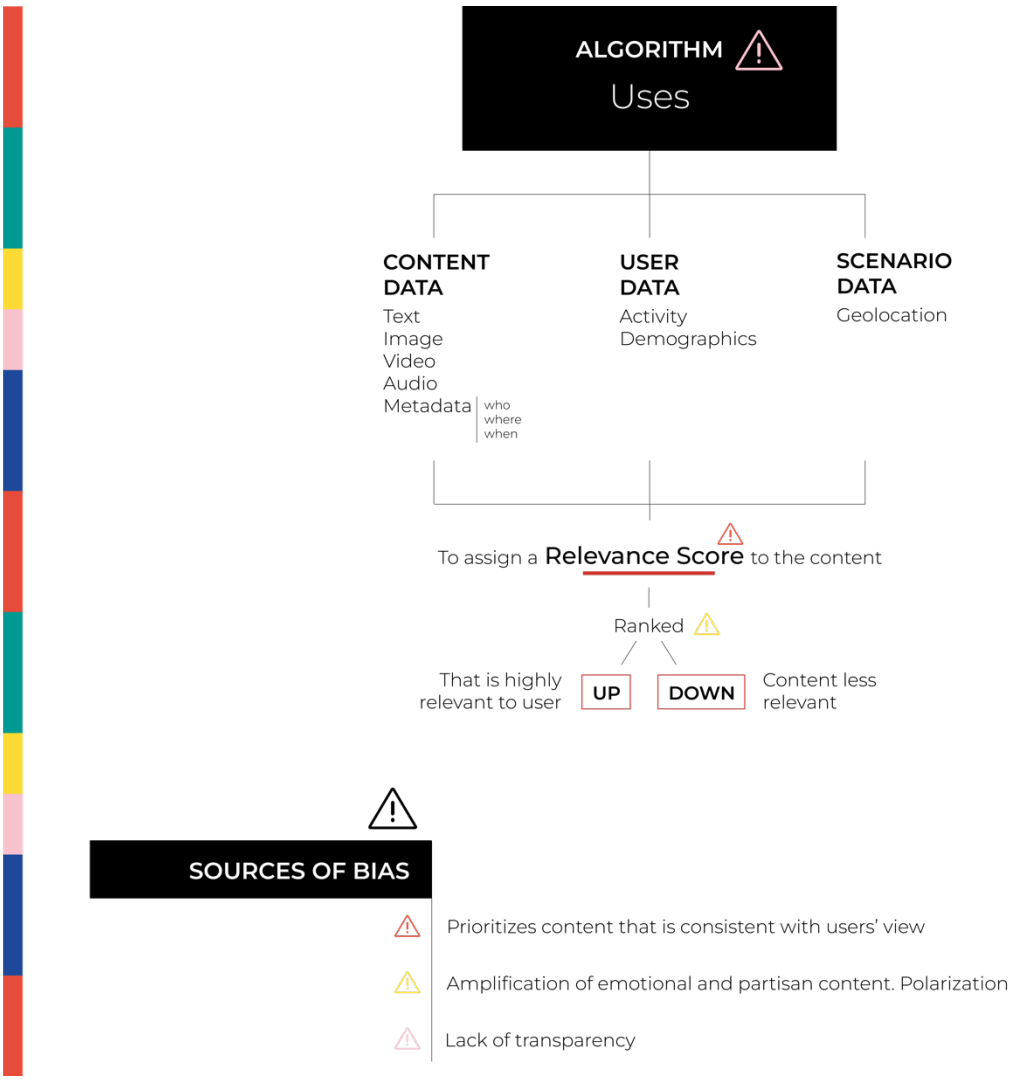

FIGURE 2: How content selection works and the different sources of bias.

2.6 ADVERTISEMENT: TARGETING AND DELIVERY

Personalisation is one of the techniques that allows social media to be powerful advertising platforms driven by profit. Personalisation allows social media platforms to attribute each user to its market, and this is the reason why online advertisement has an enormous financial success. The **targeting features** that big tech companies offer are a big source of revenue for social media companies. These targeting features allow advertisers to address very specific audiences based on users' demographics, profile information, activity on the platform, data from third parties and similarity to other potential customers (Andreou et al., 2018). The users' information that platforms exploit for targeted advertising lacks transparency and the targeting features that platforms offer vary across platforms. Generally, the targeted advertisement ecosystem involves three actors: advertisers, who decide the audience who should see the ad; ad platforms which aggregate their data and make them available for advertisers; and the users that will consume that ad.

Targeted advertising works through two main stages. **In the first stage**, the advertisers create the content of the ad (headline, text, images), choose the target audience to which they would like to show the ad and specify how much they would like to pay to have their advertisement shown (bidding). Advertisers buy space on platforms, and they place bids and the platforms display the ads of the highest bidder for a given target. Platforms “optimize” the bidding process so that advertisers only compete for people who have a larger chance of clicking on a given ad. This way, advertisers are more likely to bid for people who will click on their ads, and platforms can drain their advertising budgets more quickly. In the **second stage**, the platform delivers the advertisement to specific users through algorithmic optimisation based on a variety of factors such as budget, ad performance and relevance to the users (Ali, Sapiezynski, Bogen, et al., 2019). In this audience-ad matching process, the platform analyse all the ads placed by different advertisers in a particular interval of time, their bids and runs an auctions to determine which ones are selected. In this process, the platforms avoid showing the same advertisement to the same users so the platform might ignore bids for the same users. The platform might also include a relevance score into consideration and analyse the image and text in the ad (Ali, Sapiezynski, Bogen, et al., 2019). Once the ad is published, the platform can give advertisers information on how their ads are performing.

Targeted advertisement can also create discriminatory patterns and reinforce stereotyping.

Discrimination can happen at both stages of the advertisement process. When it comes to **targeting,** advertisers can choose targeting options that can discriminate marginalised groups. Previous work showed that advertisers were able to run ads that explicitly excluded some users from seeing it based on their race, gender, language spoken, migration status, age group, or their interest on wheelchair ramps (Angwin, Scheiber, et al., 2017; Angwin, Tobin, et al., 2017, 2017; Angwin, Varner, et al., 2017; Tobin & Merrill, 2018). Facebook targeting system was allowing to target anti-Semitic categories (Angwin, Varner, et al., 2017). **Delivery algorithms** have also been shown to discriminate users. Platforms build user interest profiles and track the performance to understand how different users interact with different ads. These data are then used to optimise the ad delivery and direct ads toward users who are most likely to engage with them. Delivery algorithms also use the characteristics of ads to optimise the delivery. For example, if an image used in an ad receives better engagement from a certain demographic, or has some gendered or racialised content, the platform's algorithm may learn the association and preferentially show the ad with that image to the subset of the targeted audience belonging to that demographic (Ali, Sapiezynski, Bogen, et al., 2019). By doing so, the platform can steer ads toward specific groups, without the advertiser being aware of it. For example, Algorithmic Watch (Kayser-Bril, 2020) advertised different jobs on Facebook and Google in different European countries (Germany, Poland, France, Spain and Switzerland). The jobs advertised were: machine learning developers, truck drivers, hairdressers, child care workers, legal counsels and nurses. Every ad used the masculine version of the job and showed a picture. The experiment used only geographical targeting as it is the only one compulsory. The truck drivers' advertisement was shown to 4,864 men and 386 women. The ad for childcare was shown to 6456 women and 258 men. Algorithmic Watch advertised the same job (Food Truck driver) with different images and text to find that Facebook appeared to use the images to decide whom to show the ad to. Political polarisation (Ali, Sapiezynski, Korolova, et al., 2019). As another example, for a job ad, the algorithm may aim to show the ad to users whose professional backgrounds best match the job ad's qualification requirements. If the targeted population of qualified individuals is skewed along demographic characteristics, the platform's algorithm may propagate this skew in its delivery. **Financial algorithmic optimisation and market effects** also affect ad delivery: some users are more expensive to advertisers than others because they are more likely to interact with the advertisement. Therefore, advertisers with lower budgets are more likely to lose the auction for those users and even if an advertiser might not explicitly choose to exclude a user, the ad delivery process might, because of the higher demand of that group. Lambrecht & Tucker (2018) for example showed that gender-neutral ads on job opportunities in STEM were seen by more men than women on Twitter, Facebook, and Instagram. When women were shown

the advertisement, they were more likely to click on it than men, but women were a prized demographic, so they were more expensive to show ads to.

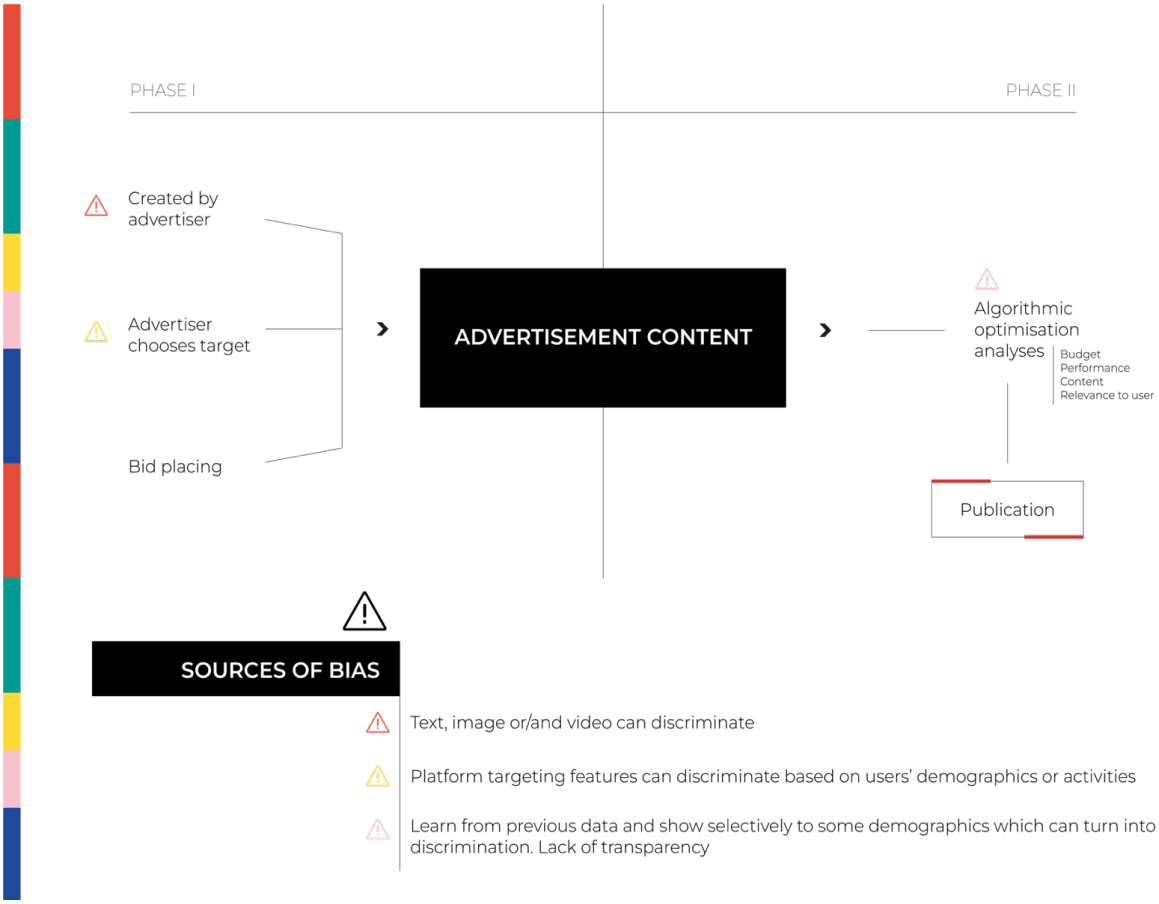

FIGURE 3: How targeted advertisement works and the different sources of bias.

2.7 Conclusion

Section 2 presented how social media content moderator, shadowban, personalisation of content and targeted advertising works, and how they affect the representation of marginalised groups. It discussed how social media dynamics can contribute to censorship, amplification and manipulation of online information. Between the creation and publication of the content there are different stages of human and algorithmic filtering that can contribute to the amplification of stereotypes and to silence the voices of marginalised groups which reproduce unequal social structures. The illustration below summarizes these stages. Section 3 and 4 will explore how social media affects the experience and representation of migrants and refugees.

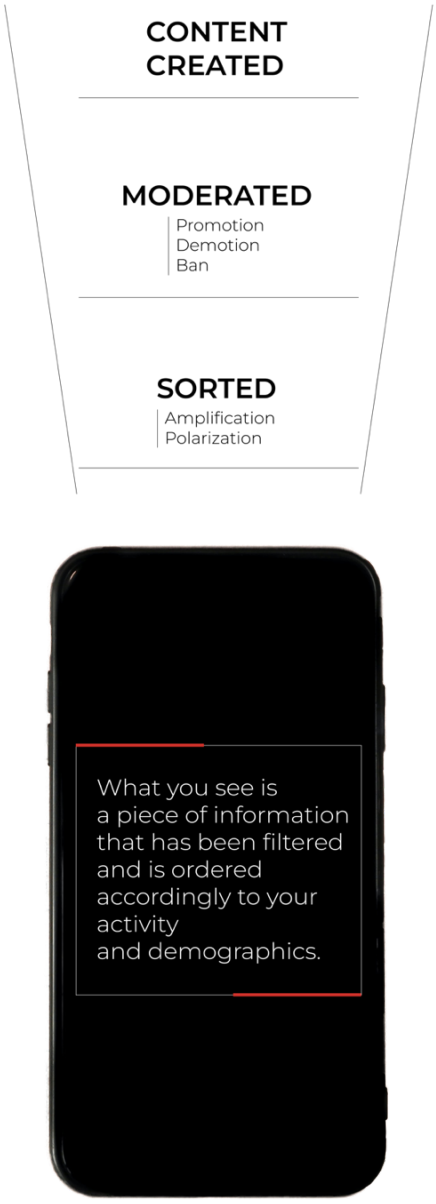

Figure 4: Summary of how the content is filtered and selected.

3. THE REPRESENTATION OF MIGRANTS AND REFUGEES IN SOCIAL MEDIA

In recent years, migration has been dominating media and political discourse in Europe. People with migrant backgrounds are generally absent from this discourse, which is particularly problematic as it disables the community to defend itself from attacks and make migrants authors of their own stories. Generally, research has focused on mapping discourse in traditional media or conventional channels of communication, mainly depicting migrants under a negative light. Migrants are often represented as delinquents, criminal or economic and cultural burden for the society (Eberl et al., 2018). Migration in the media is constructed in ways that dehumanize migrants and position them as outgroups (van Klingeren et al., 2015), as threat to cultural norms or economic wellbeing and danger to social norms (Murray & Marx, 2013). The news stories tend to focus on crime, public unrest, cultural misunderstandings, social problems, and economic costs (Rosina, 2022). Migrants are not seen as complex individuals with complex stories but othered and portrayed either as victims or perpetrators. Researchers have shown that the way migrants are depicted in the media impact individual and societal attitudes toward immigration (van Klingeren et al., 2015).

In the last decades, the way people access and share information changed because of the adoption of social media. Mainstream-media content still plays a major role in shaping discourse about events and social groups, but social media's participatory affordances allow for narratives to be perpetuated, challenged, and injected with new perspectives by social media users. This section presents how migrants use social media in their migration process, its benefits and risks, and how social media represents migrants and refugees. Section 3.1 reveals that social media is increasingly used by migrants in their migration process, but their voice is often silenced and threatened by digital surveillance, which, in addition to the social media dynamics affecting vulnerable populations described in section 2, discourage migrants to speak openly and to reveal their stories and identities online. Section 3.2 describes how migrants are represented on social media and shows how migrants and refugees' representation is related to stereotyping and issues of hate speech, incitement to violence and misinformation.

3.1 MIGRATION AND SOCIAL MEDIA

In February 2022, two TikTok videos from two women drew media attention. The videos were documenting their journey to Europe on a boat with other migrants (Ebel, 2022). These videos are just two examples of the many videos available on TikTok under the hashtag "harka" or "haraga", which are both used in North Africa to refer to Mediterranean crossing (Joles, 2022). Technology is

transforming trends of migrations and migrants are increasingly relying on social media to organise and document their movements.

> "For refugees seeking to reach Europe, the digital infrastructure is as important as the physical infrastructures of roads, railways, sea crossings and the borders controlling the free movement of people. It comprises a multitude of technologies and sources: mobile apps, websites, messaging and phone calling platforms, social media, translation services, and more."(Gillespie et al., 2016)

Another example of how social media are changing migration is the case of Rahf Mohamed, a Saudi teenager trying to flee from her abusive family. Rahf Mohamed used Twitter to document her experiences of psychological and physical abuse. A few hours after her first tweet, users created the hashtag #SaveRahaf which was shared more than half a million times and brought the case to the attention of governments, media and the UNHCR. As a result, Rahf Mohamed was granted asylum by the Canadian government (BBC, 2019). Smartphones and their apps are radically changing migration processes. Because of this, researchers have recently started to discuss digital migration as a new and expanding field of study (Leurs & Smets, 2018).

The transformation that social media is affording to migration has both favourable and unfavourable consequences. Social media empower migrants and refugees by strengthening their social networks, helping them to find information, connect with humanitarian associations, take decisions, and document their journeys and situations. At the same time, social media can contribute to expose migrants to misinformation and to the danger of getting in touch with smugglers, who are increasingly using these platforms to advertise and coordinate their services (Europol, 2022a). On the other hand, governments across sectors are resorting to social media intelligence[4], which is used to manage the migration flows and verify migrants' identities and their narratives. This makes social media a tool of digital surveillance that undermine human rights and discourage undocumented migrants and

[4] Social media intelligence comprises a series of techniques and technologies that allow companies and governments to monitor social media networking platforms.

refugees to share their stories and identities online. For these reasons, migrants and refugees tend to hide their identities via avatars and pseudonyms (Leung, 2010). In Europe, governmental border controls and migration managements increasingly rely on digital technologies to surveil and detect migrants: they scrape social media data for predictive policing and look at social media profiles of migrants and asylum seekers to corroborate the information they provide to the authorities (Bolhuis & van Wijk, 2021; Brekke & Balke Staver, 2019). For example, The European Asylum Support Office (EASO) monitored refugee networks to detect new routes and find smugglers, a practice that stopped in 2019 after the European Data Protection Supervisor (EDPS) imposed a temporary ban, saying EASO had no legal basis for monitoring refugee routes on social media (Privacy International, 2017). Frontex was planning a similar programme of social media scraping to detect migrants routes and activities (Kilpatrick & Jones, 2022). Other international organizations are turning to big data to predict population movements, and laws in countries like Austria, Denmark, Germany, Norway, the UK and Belgium, allow for the seizure of mobile phones from asylum or migration applicants from which data is then extracted and used as part of asylum procedures (Biselli & Beckmann, 2020; Meaker, 2018). Twitter has also been also used to analyse migration movement using both twitter text with relevant information on migrant movements and geolocation (Urchs et al., 2019). These practices are a serious interference with migrants' privacy and tend to criminalise the migrants' community (Rosina, 2022).

3.1.2 Communication, information and surveillance

For undocumented migrants and refugees' accessibility to information is vital. Because of this, social media has become an essential tool that provides them access to information that supports them prior to, during and after their migration journey. Social media are important **sources of information** that affect migrants' intentions to migrate and decisions on *how*, *whether* and *where* to migrate (Dekker et al., 2018).

> Social media platforms make prospective migrants more informed about possibilities to migrate, bureaucratic processes and destinations where to settle, enabling them to develop their migration strategies.

Researchers showed that sub-Saharian migrants used smartphones and social media to access online information before, during their travel and when they arrived at their destination country. Smartphones and social media affected decisions on their migration routes and final destination while

helping migrants to share information with each other (Ennaji & Bignami, 2019). Ennaji & Bignami (2019) found that smartphones and the accessibility to information impacted people's intention to migrate. Smartphones made the migration process smoother as migrants could use google maps to reach their destinations and social media to connect and stay in touch with family, friends (Laub, 2019) and influential figures in their social media network such as activists, NGOs, investigative journalists, political commentators, public intellectuals and participants in controversial debates (Gillespie et al., 2016). Social media are also used during the migration journey to ask for help, in case of loss or need. They are used to establish new social ties, to receive advice from people who already live abroad and to communicate with people met in the migration journey while crossing borders (Schmidle, 2015).

For example, research focused on Syrian refugees pointed out that they used Facebook groups to discuss news on their home countries, information about asylum processes, asked questions related to refugees' experiences, and shared news about routes and journey experiences (Schmidle, 2015). The Facebook group "Asylum and Immigration without Smugglers" founded by the Syrian refugee Abu Amar in 2013 is one of those places where refugees find information about borders, routes, weather conditions and places where to stay (Schmidle, 2015). Furthermore, experts pointed to TikTok as a new entry point for young people into the world of irregular migration where recommendation algorithms might suggest migration content to young people that are not even searching for it (Joles, 2022). As the Rest of World report shows, TikTok offers an ecosystem of content that is made by and for migrants that documents journeys. Generally, the videos imply that the passage was irregular but they do not say it explicitly as content referring to smuggling services would be subjected to ban in Facebook or Instagram (Joles, 2022). However, the moderation of this content is difficult. From a linguistic perspective, social media might struggle to understand the content of regional accents that might refer to irregular migration. Language is also used to keep the content from being flagged and banned.

Ennaji & Bignami (2019) argued that because of the access to information that social media afford, some migrants are making their journeys without smugglers. However, easy access to new social networks can facilitate exposure to misinformation and their contact with smugglers who have expanded their use of social media to offer their services (Europol, 2022b). For example, social media targeting features enable migrants to easily advertise their services to individual profiles, groups and pages. Smugglers can share on social media a range of information on prices, contacts, testimonials from previous clients that might persuade people to use their services (Roberts, 2017). Migrants - especially asylum migrants - deal with information precarity when using social media (Dekker et al., 2018). Information precarity refers to a state in which migrants access information is insecure and

unstable, leading to potential threats to their wellbeing, and leaving them vulnerable tomisinformation, stereotyping and rumors that can affect their economic and social capital (Wall et al., 2017). This precarity refers both to information access and information content. In terms of access, refugees often struggle to have stable internet connection (UNHCR, 2016). In terms of content, there are issues with finding the information needed and determining the trustworthiness of that piece of information. This can expose migrants to misinformation. Gillespie et al., (2016) showed that to deal with this precarity, refugees gathered information from closed groups, as they were judged as more trustworthy and that they deemed as more trustworthy information coming from existing social ties and based on personal experience (Dekker et al., 2018). This research reported that migrants were making little use of mainstream national and international media sources as here was a fundamental mistrust of western individuals and organisation when it came to the engagement on Facebook groups. Gillespie et al. (2016) also showed that relationships among migrant on social media were shaped by kinship and friendship but also by pragmatic and ideological factors. In line with the research reported in Section 2, Gillespie et al. (2016) highlighted that the spaces of social media discussion and debate among refugees tended to be polarised and politicised and that influential figures directed the flow of engagement and information within a network that reinforced and maintained the networks as insular ideological enclaves.

A qualitative study with refugees from Syria, Eritrea and Afghanistan also pointed out the role of social media for the settlement of the refuges in the new context (Alencar, 2018). Social media was important to acquire language, to bond with other refugees, to establish contacts with locals and to bridge social capital. Social media can provide migrants with general information about rights, citizenship, and local migrant support services. The differential role of social media compared to other Internet-based applications relies on the development of migrants' social networks and the possibility of users to consume, produce and share content and opinions within and across networks (Sawyer & Chen, 2012). Furthermore, one participant in Alencar (2018) study pointed out that social media could help to change negative stereotypes toward refugees through Facebook pages and groups including information about the habits, languages, and traditions. Section 3.2 will review how migrants and refugees are framed in social media.

3.2 How are migrants and refugees framed in social media?

A review on media representation of migrants published in 2018 (Guidry et al., 2018) emphasized that the majority of researchers' work has focused on national media systems and print media outlets while social media and user generated content are neglected in research on migrants' representation. Studying social media is important to be able to map the discourse on migration in broader terms and understand the issues related to migrants' online representation. This section will outline how the representation of migration in social media is related to issues of hate speech, incitement to violence and disinformation and it will review recent studies on migrants' representation in major social media such as Twitter, Facebook, Instagram, YouTube.

Hate speech (Arcilla Calderòn et al., 2020; Ekman, 2019; UN Migration, 2019) refers to the use of discriminatory and derogatory language on the basis of migrants' religion, nationality, race, descent, gender and other identities. Examples of hate speech are the Instagram and Facebook hashtag #stopislam (Civila et al., 2020) which shows fake material inciting other users to engage in hate speech against migrants. This hashtag is also sometimes used by people that create pro-immigrant counter narratives. **Incitement to violence** (European Union Agency for Fundamental Rights, 2016) refers to speech that trigger discrimination, violence toward migrants and might also lead to crime. In this context, a study conducted in Germany between 2015-2017 showed that during times of higher anti-immigrant sentiments, areas with higher numbers of Facebook users had an increase of up to 50% of anti-refugee incidents than the national average (Tidey, 2018). **Disinformation, fake news and conspiracy theories** (Schafer & Schadauer, 2018; Wright et al., 2021) also mislead users' understanding of migration issues and migrants' experience, often for political ends and to raise anti-migrants attitudes. One research on online disinformation showed that most stories in the media present migrants as a health threat, economic threat or criminal threat (Cerase & Santoro, 2018). As a health threat, hoaxes rely on the representation of immigrants and refugees as unclean and poor hygiene habits which lead to allegations of being the carrier of diseases. For example, Facebook users were found to spread content about an Ebola epidemic in Lampedusa, an arrival point for migrants, that needed to be debunked by the government (Cerase & Santoro, 2018).

Hoaxes have also been found to be about immigrants escaping the quarantine or not observing the lockdown rules (Klein, 2020). Avaaz, (2019) revealed that ahead of the 2019 Spanish elections disinformation and hateful content reached up to 9.6M potential voters via WhatsApp. 14% of the content was spreading anti-immigration fake news. Some of these news regarded the representation of migrants as an economic threat. Messages were about immigrants receiving 1,800 euros per month in social welfare, or receiving more money than a pensioner evicted from his house by the Spanish

state. 25% of the WhatsApp content was also found to be openly racist. Often, the hoaxes use old pictures that are recontextualized to justify the fake news described on text.

Online disinformation tends to rely on stereotypes and societal bias that are reproduced and reinforced on the network online. The section below will show that social media representation of migrants and refugees is associated with both positive and negative tones and is highly politicized, framed in pro-migration and anti-migration narratives that refer to them either as vulnerable groups or as a threat.

> Disinformation, hate speech and incitement to violence pose a real threat to minority groups that are object of discrimination.

3.2.1 INSTAGRAM

De Rosa et al., (2020) analyzed 456 manually selected Instagram photos with textual elements referring to migratory issues. This research reported a dichotomous discourse about immigration which opposed positive social representations of inclusive practices (e.g. welcome refugees) and negative social representations of exclusionary practices (e.g. closed ports). This polarized discourse revealed representations of migrants as vulnerable groups opposed to the portrayal of migrants as dangerous invaders. While the first was linked to a tendency of solidarity, the second was more focused on legal, economic and ethnic aspects for political propaganda.

Another study (Guidry et al., 2018) focused on refugees randomly sampled 750 Instagram posts and 750 Pinterest posts published between February and April 2016. The authors distinguished between thematic framing -where the story presents broader societal and political patterns- and episodic framing – where the story focuses on one case. Their results showed that both on Pinterest and Instagram security concerns were likely to be framed thematically while humanitarian concerns were more likely to be framed episodically. Instagram revealed more episodic framing while Pinterest contained more thematic framing with more expression of fear of refugees and perceptions of refugees as dangerous. On the other hand, Instagram displayed more humanitarian-concern expressions. On both platforms, posts were related to fear, terror, anxiety both from a humanitarian and a security perspective. Authors have also used Instagram to explore how migrants were

represented in Idomeni (Radojevic et al., 2020). According to this research, the frames were primarily concerned with human impact, transnational conflict constellations, ethics, and responsibility.

Migration is a topic that is highly politicised. Jaramillo-Dent et al., (2022) explored right wing immigration narratives in Spain analyzing 832 Instagram stories related to immigration published by a Spanish right-wing party (Vox). In this study, researchers coded the Instagram stories and the results showed that the prevalent depiction of migrants was those of an unidentified, black, and male migrant which is part of a group. Migrants were represented as perpetrators of violence and fraud to build persuasive ideological narratives.

3.2.2 YOUTUBE

Researchers (Lee & Nerghes, 2018) have compared comments to the two most popular YouTube videos using the "refugee crisis" and "migrant crisis" labels. The authors discovered distinct topics with different valences for each framing and that comments under video framed as "migrant crisis" were less varied. Comments under the "migrant crisis" videos had fewer positive topics as compared with comments under the video framed as "refugee crisis". Less negative topics were found in comments under the "refugee crisis" video, but overall sentiments related to how migrants and refugees were framed were negative in both cases.

Another study on YouTube analysed the content of racist and xenophobic comments written in Spanish against migrants and refugees (Latorre & Amores, 2021). This study showed that hate speech followed right wing political ideologies, and represented migration as a burden (e.g., Spaniard first) and economic (e.g. immigrants take our jobs), cultural (e.g. close borders to save culture and values) and security threat.

Aguirre & Domahidi (2021) found that comments to YouTube videos on the Venezuelan "refugee crisis" were both offensive (32%) and hateful (20%) and hateful comments had higher number of likes than offensive ones. The hateful comments were racist, xenophobic and sexist while offensive comments used derogatory terms, urged immigrants to leave the country and offended people with left wing political views. Generally, this problematic content was depicting migrants under stereotypes of people that are lazy and that engage in irregular activities. This study also pointed out a social network behind the structure of problematic content. Problematic content came from a small percentage of active users that were interconnected with each other and were responsible for 40% of the negative content.

Spörlein & Schlueter (2021) indicated that the use of ethnic insults targeting ethnic minorities in the YouTube comment section appear contagious, a dynamic that was also found in other research fields

(Kwon & Gruzd, 2017). The authors analysed ethnic insults in the comment sections of YouTube videos from the four most popular German political talk shows before, during, and after the height of the European “immigration crisis”. They found that a larger presence of ethnic insults in preceding comments increased the prevalence of insulting comments by 2 percentage points, which increased to 7 percentage points in the aftermath of violent incidents. However, when the dataset was restricted only to frequent commenters, this relationship was not found. They also found that ethnic insulting commenting became more viral in periods of violent attacks committed by minority members, but not among frequent commenters.

3.2.3 TWITTER

Studies addressing the topic of the refugee and migrants on Twitter expose discourses from solidarity and social justice to xenophobia (Gualda & Rebollo, 2016) and racism (Siapera et al., 2018). They showed a dichotomy between the "deserving" refugee versus the "undeserving" migrant (Nerghes & Lee, 2018) and different frames of refugees and migrants portrayed as vulnerable population or threat. Comparative studies on the representation of migrants on mainstream media and Twitter (Nerghes & Lee, 2019) argued that social media and mainstream media interact in the creation of narratives around migrants, with Twitter playing an important role for digital activism. Nerghes & Lee (2019) showed that tweets about migrants and refugees displayed more sympathetic tones and introduced new themes into the discussion that pushed the debate into new directions compared to the narratives created by news media whose topics were more focused on geopolitics. Twitter users created an alternative narrative through solicitations of sympathy and calls-to-action while mainstream media politicised the content on migrants and refugees. The solicitations of sympathy on Twitter included placing oneself in the shoes of the refugees, promoting a migrant success story, and promoting empowered figures. On the other hand Siapera et al. (2018) showed dehumanising and politicised Twitter narratives revolving around key events that were widely publicised through mainstream media. These authors found that migrants' stories were politicized from a far-right perspective and framed migrants as terrorists or rapists to mobilize security and safety preoccupations. There was also a humanitarian frame, which revolved around human rights, and was created by humanitarian organizations, activists, and some mainstream media. The evidence provided by this study showed processes of othering and dehumanising migrants on Twitter.

Other studies focused on tweets produced in national contexts. Research carried out in Sweden (Yantaseva, 2020) showed that participatory media were more focused on the social effects of migration and were found to incorporate a variety of potentially prejudiced frames. Results of this research revealed that 7 out of 10 frames on both Twitter and an online forum displayed a negative

sentiment. In social media specifically, the research identified a “humanitarian crisis” frame that discussed migration in the context of human suffering and assistance to refugees, including terms that were related to sympathy (home, help, responsibility, war). At the same time, a variety of frames focused on racial, ethnic, and religious differences between people or presented negative aspects of migration. Some of the frames included explicitly racist or negatively loaded terms. Other authors analysed political hate speech and found that tweets contained offenses, incitements to hate and violent speech (Arcilla Calderòn et al., 2020).

Gualda & Rebollo, (2016) showed that the discourse on refugees in Europe presented both a positive and a negative light. On one hand, they displayed solidarity speech supporting refugees and claiming social justice highlighting Europe’s responsibility toward children and people’s vulnerabilities. On the other hand, they recorded negative speech focused on religious identities and using metaphors like "invasion" or talking about terrorism. Other work analysing the sentiments of tweets revealed the prevalence of negative sentiments in messages related to refugees (Ladner et al., 2019). Kreis (2017) analysed tweets with the #refugeesnotwelcome hashtag. The study revealed that the hashtag relied on a rhetoric of inclusion and exclusion that depicted refugees as unwanted and criminal outsiders, which overlapped with right wing national conservative anti-immigrant narratives. Nerghes & Lee (2018) focused on tweets on the aftermath of Alan Kurdi drowning and showed that tweets were focused on "refugees" hashtags and were more positive in the tone. Furthermore, they detected that popular Twitter users, as well as popular tweets, were characterized by less emotional intensity and slightly less positivity in the debate.

3.2.4 FACEBOOK

Ekman, (2019) analysed how immigrants and refugees are discursively constructed in the Facebook group ‘SufS’ (Stand up for Sweden), and how racism was overtly and covertly expressed in user comments in the group. Drawing on both quantitative and qualitative content analysis, the article emphasized the particular role of emotions in shaping racist discourse. The authors showed that racist opinions and attitudes became normalized through the recontextualization of mainstream news covering refugees and migrants, and that the affective character of public comments triggered racist attitudes. Content of comments was positioning Swedish citizens as victims of cultural threat and involved racist and dehumanising comments under news regarding criminal activities.

Capozzi et al., (2020) examined political and social targeted advertising concerning immigration in Italy. They found that advertised content related to immigration increased during election times and that top spenders’ parties were the ones promoting anti-immigration policies. The framing varied: right wing parties' advertisement focused on “clandestines” (i.e., undocumented immigrants) and

referred to rescue ships operations. They emphasized themes of sovereignty by referring to law, territory, nation, and citizenship. Instead, leftist parties mentioned words related to rights and duties such as “law” and “rights”. This research also showed that the final reach of the advertisement campaign was characterised mostly by male and older people and that distribution changed depending on the party, indicating targeting.

Heidenreich et al. (2020) analysed political Facebook accounts across six European country (Spain, UK, Germany, Austria, Sweden, and Poland) to show that migration was a more prominent topic in receiving countries -countries in which there is a positive net migration- than in countries where the net migration was neutral or negative. They showed that migration was more prominent in the Facebook posts of more ideologically extreme parties. Similarly, the more ideologically extreme the party, the more negative the sentiment of its migration-related status posts.

3.2.5 Search Engines

A recent work (Urman et al., 2022) audited the representation of migrants in image web search results and showed that search engines reproduce social biases by proposing content that aligns with societal stereotypes. This content reproduced gender and ethnic bias of different groups of migrants. The authors collected image web search results related to various terms referring to migrants such as expats, immigrants, and refugees. They indicated that migrants and refugees’ representations tended to be highly racialized and that female migrants and migrants at work tended to be under-represented for the migrant query. Women and children were overrepresented for the refugee query.

3.3 Conclusion

Section 3 uncovered how unbalanced structures of power are maintained by social media content with a focus on the migrant and refugee population. The section showed that migrants and refugees are not fairly represented on social media. Migrants and refugees are often depicted under a negative light, stereotyped and the object of hate speech, incitement to violence and misinformation. Social media content is highly politicized with narratives built for them and not by them. This reduces the complexities of their stories and dehumanises them. Beyond being victims of social media algorithmic systems, refugees and migrants are also discriminated by the content that is produced in these platforms. This needs to be taken into account when shaping regulations of the digital space: social media needs to be kept accountable of the way they manage particular and contextual risks related to these communities. Additionally, the section presented how social media are changing migration processes affording people a better communication and information but exposing them to new forms of digital surveillance. In the migration context, digital surveillance poses a real threat to people's rights of free movement, privacy and non-discrimination and further silence migrants and refugees' voice, who already struggle to be fairly represented as complex human beings.

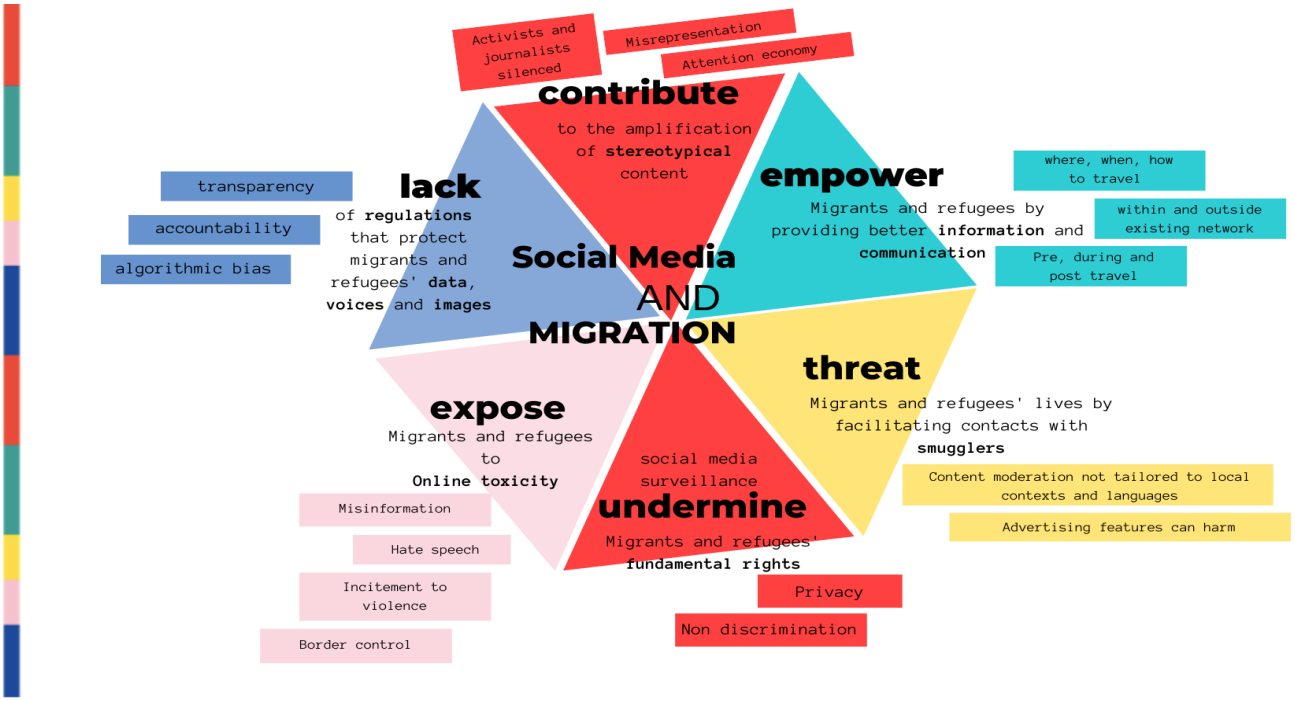

FIGURE 5: Social media and migration.

4. GENERAL CONCLUSION

In the last decades, social media have become ubiquitous in the way people communicate, consume information and network with others. The big tech companies behind social media are huge economic empires who have the control over the information users are exposed to, using opaque algorithms informed by users' data that keeps them engaged and accommodating existing unequal power structures. There is a lack of transparency, accountability and regulatory frameworks that alleviates social media companies from taking responsibility for their wrongdoings. The socio-technical dynamics that characterise social media platforms create and allow for new forms of stereotyping, discrimination, silencing of voices and manipulation of reality that affect all but hit harder the representation of marginalised social groups.

Migrants and refugees are among those who are affected by the new forms of discrimination social media dynamics create. These new forms of discrimination are determined by platform business models, lack of transparency and accountability, algorithmic bias and ideological views that stereotype refugees and migrants and make them invisible on social media. In the context of "Re-framing migrants", this report identified key social media dynamics that contribute to discrimination, stereotyping and silencing of voices with a focus on the migrant population.

The report started by explaining how content moderation, shadowbanning, content selection and targeted advertisement silence voices and create issues of amplification, discrimination and manipulation of reality that harshly affect the representation of minorities and marginalised groups. It discussed how social media is changing migration by allowing for better communication and information but exposing migrants and refugees to surveillance threats. Finally, it showed how migrants' and refugees' representation on social media is related to stereotyping and issues of hate speech, incitement to violence and misinformation.

Examining how social media misrepresent migrants and silence their voices enables to challenge and change the social media socio-technical infrastructure by:

- identifying techniques and tools to subvert the social media system,
- planing strategies to address the power asymmetries that characterise the social media information space,
- designing strategies and actions to amplify those voices that are now excluded from the debate,

- shaping regulations that contextually consider the risks to which social media expose specific social groups.

REFERENCE LIST

Aguirre, L., & Domahidi, E. (2021). Problematic Content in Spanish Language Comments in YouTube Videos about Venezuelan Refugees and Migrants. *Journal of Quantitative Description*, *1*. https://doi.org/10.51685/jqd.2021.022

Alencar, A. (2018). Refugee integration and social media: A local and experiential perspective. *Information, Communication & Society*, *21*(11), 1588–1603. https://doi.org/10.1080/1369118X.2017.1340500

Ali, M., Sapiezynski, P., Bogen, M., Korolova, A., Mislove, A., & Rieke, A. (2019). *Discrimination through optimization: How Facebook's Ad delivery can lead to biased outcomes.* 1–30.

Ali, M., Sapiezynski, P., Korolova, A., Mislove, A., & Rieke, A. (2019). *Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging*. https://doi.org/10.48550/ARXIV.1912.04255

Andreou, A., Venkatadri, G., Goga, O., Gummadi, K., Loiseau, P., & Mislove, A. (2018). *Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations*. *1–15*. https://hal.archives-ouvertes.fr/hal-01955309/file/Andreou-etal_FacebookAdExplanations_NDSS2018.pdf

Angwin, J., Scheiber, N., & Tobin, A. (2017). *Dozens of Companies Are Using Facebook to Exclude Older Workers From Job Ads*. https://www.propublica.org/article/facebook-ads-age-discrimination-targeting

Angwin, J., Tobin, A., & Varner, M. (2017). Facebook (Still) Letting Housing Advertisers Exclude Users by Race. *ProPublica*. https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin

Angwin, J., Varner, M., & Tobin, A. (2017). *Facebook Enabled Advertisers to Reach 'Jew Haters.'* https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters

Arcilla Calderòn, C., De la Vega, G., & Blanco Herrero, D. (2020). Topic Modeling and Characterization of Hate Speech against Immigrants on Twitter around the Emergence of a Far-Right Party in Spain. *Social Sciences*, *9*(11). https://doi.org/10.3390/socsci9110188

Are, C. (2021). The Shadowban Cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*, 1–18. https://doi.org/10.1080/14680777.2021.1928259

Avaaz. (2019). *Whatsapp Social Media's Dark Web. How the messaging service is being flooded with lies and hate ahead of the spanish elections.* https://avaazimages.avaaz.org/Avaaz_SpanishWhatsApp_FINAL.pdf

Bandy, J. (2021). *Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits* (arXiv:2102.04256). arXiv. http://arxiv.org/abs/2102.04256

BBC. (2019). *Rahf Mohammed: Saudi teen says women "treated like slaves."* https://www.bbc.com/news/world-us-canada-46873796

Biddle, S., Ribeiro, P. V., & Dias, T. (2020). Invisible censorship. TikTok Told Moderators to Suppress Posts by "Ugly" People and the Poor to Attract New Users. *The Intercept*. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/

Biselli, A., & Beckmann, L. (2020). *Invading Refugees' Phones: Digital Forms of Migration Control in Germany and Europe*. Gesellschaft für Freiheitsrechte e.V. https://legacy.freiheitsrechte.org/home/wp-content/uploads/2020/02/Study_Invading-Refugees-Phones_Digital-Forms-of-Migration-Control.pdf

Bolhuis, M. P., & van Wijk, J. (2021). Seeking Asylum in the Digital Era: Social-Media and Mobile-Device Vetting in Asylum Procedures in Five European countries. *Journal of Refugee Studies*, *34*(2), 1595–1617. https://doi.org/10.1093/jrs/feaa029

Brekke, J.-P., & Balke Staver, A. (2019). Social media screening: Noway's asylum system. *Forced Migration Review*, *61*, 9–11.

Brems, C., Temmerman, M., Graham, T., & Broersma, M. (2017). Personal Branding on Twitter: How employed and freelance journalists stage themselves on social media. *Digital Journalism*, *5*(4), 443–459. https://doi.org/10.1080/21670811.2016.1176534

Bruno, N. (2011). Tweet first , verify later? How real-time information is changing the coverage of worldwide crisis events. *Reuters Institute for the Study of Journalism*.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, 1:15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Byers, D., & Abbruzzese, J. (2020). Facebook employees go public with disagreement over Zuckerberg's handling of Trump. *NBC News*. https://www.nbcnews.com/tech/tech-news/facebook-employees-go-public-disagreement-over-zuckerberg-s-handling-trump-n1220961

Capozzi, A., G. D, M. F., Mejova, Y., Monti, C., Panisson, A., & Paolotti, D. (2020). Facebook Ads: Politics of Migration in Italy. Inpp. 43-57). Springer, Cham. *International Conference on Social Informatics*, *43–57*. https://link.springer.com/chapter/10.1007/978-3-030-60975-7_4

Cen, S., & Shah, D. (2021). Regulating algorithmic filtering on social media. *Advances in Neural Information Processing Systems*, *34*. https://papers.nips.cc/paper/2021/file/38b4f06e27fd4f6fdcceabc6f5c068ea-Paper.pdf

Cerase, A., & Santoro, C. (2018). From racial hoaxes to media hypes: Fake news' real consequences. In *From Media Hype to Twitter Storm*. Amsterdam University Press. https://doi.org/10.2307/j.ctt21215m0

Chowdhury, R. (2021). *Sharing learnings about our image cropping algorithm*. Https://Blog.Twitter.Com/. https://blog.twitter.com/engineering/en_us/topics/insights/2021/sharing-learnings-about-our-image-cropping-algorithm

Civila, S., Romero-Rodríguez, L. M., & Civila, A. (2020). The Demonization of Islam through Social Media: A Case Study of #Stopislam in Instagram. *Publications*, *8*(4), 52. https://doi.org/10.3390/publications8040052

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showd bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *ArXiv Preprint*, 1905.12516.

De Rosa, A. S., Bocci, E., Nubola, A., & Salvati, M. (2020). The Polarized Social Representations of immigration through the photographic lens of INSTAGRAM. *Psychology Hub*, *V. 37*, 5-22 Paginazione. https://doi.org/10.13133/2724-2943/17227

Dekker, R., Engbersen, G., & Faber, M. (2016). The Use of Online Media in Migration Networks: The Use of Online Media in Migration Networks. *Population, Space and Place*, *22*(6), 539–551. https://doi.org/10.1002/psp.1938

Dekker, R., Engbersen, G., Klaver, J., & Vonk, H. (2018). Smart Refugees: How Syrian Asylum Migrants Use Social Media Information in Migration Decision-Making. *Social Media + Society*, *4*(1), 205630511876443. https://doi.org/10.1177/2056305118764439

Dias, A., McGregor, J., Day, L., triple J Hack, & Four Corners. (2021). The TikTok spiral. *ABC News*. https://www.abc.net.au/news/2021-07-26/tiktok-algorithm-dangerous-eating-disorder-content-censorship/100277134

Dickey, M. R. (2020). *Twitter and Zoom's algorithmic bias issues*. https://techcrunch.com/2020/09/21/twitter-and-zoom-algorithmic-bias-issues/

Duarte, N., Llanso, E., & Loup, A. (2017). *Mixed Messages? The limits of automated social media content analysis*. Center for Democracy & Technology. https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf

Ebel, F. (2022). *Tunisian women's posts glamorize risky migrant crossings*. https://apnews.com/article/coronavirus-pandemic-business-health-africa-europe-0ab58fb921ca234561d09341f9703460?utm_source=Twitter&utm_campaign=SocialFlow&utm_medium=AP

Eberl, J.-M., Meltzer, C. E., Heidenreich, T., Herrero, B., Theorin, N., Lind, F., Berganza, R., Boomgaarden, H. G., Schemer, C., & Strömbäck, J. (2018). The European media discourse on immigration and its effects: A literature review. *Annals of the International Communication Association*, *42*(3), 207–223. https://doi.org/10.1080/23808985.2018.1497452

Eddy, S. L., & Hogan, K. A. (2014). Getting Under the Hood: How and for Whom Does Increasing Course Structure Work? *CBE—Life Sciences Education*, *13*(3), 453–468. https://doi.org/10.1187/cbe.14-03-0050

Ekman, M. (2019). Anti-immigration and racist discourse in social media. *European Journal of Communication*, *34*(6), 606–618. https://doi.org/10.1177/0267323119886151

Ennaji, M., & Bignami, Fi. (2019). Logistical tools for refugees and undocumented migrants: Smartphones and social media in the city of Fès. *Work Organisation, Labour & Globalisation*. https://www.scienceopen.com/hosted-document?doi=10.13169/workorgalaboglob.13.1.0062

European Union Agency for Fundamental Rights. (2016). *Incitement in media content and political discourse in EU member States*. https://ec.europa.eu/information_society/newsroom/image/document/2016-47/fra_media_and_incitement_paper_19752.pdf

Europol. (2022a). *Migrant Smuggling Centre 6th Annual Report*. https://www.europol.europa.eu/cms/sites/default/files/documents/EMSC%206%20th%20Annual%20Report.pdf

Europol. (2022b, February 23). *Migrant smugglers and human traffickers: More digital and highly adaptable*. https://www.europol.europa.eu/media-press/newsroom/news/migrant-smugglers-and-human-traffickers-more-digital-and-highly-adaptable

Gayle, D. (2021). Facebook aware of Instagram's harmful effect on teenage girls, leak reveals. *The Guardian*. https://www.theguardian.com/technology/2021/sep/14/facebook-aware-instagram-harmful-effect-teenage-girls-leak-reveals

Gillespie, M., Ampofo, L., Cheesman, M., Faith, B., Iliadou, E., Issa, A., Osseiran, S., & Skleparis, D. (2016). *Mapping Refugee Media Journeys Smartphones and Social Media Networks*. The Open University, France Médias de Monde. https://www.open.ac.uk/ccig/sites/www.open.ac.uk.ccig/files/Mapping%20Refugee%20Media%20Journeys%2016%20May%20FIN%20MG_0.pdf

Gillespie, T. (2018). *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Global Witness. (2022). *Facebook approves adverts containing hate speech inciting violence and genocide against the Rohingya*. https://www.globalwitness.org/en/campaigns/digital-threats/rohingya-facebook-hate-speech/

Gualda, E., & Rebollo, C. (2016). The refugee crisis on Twitter: A diversity of discourses at European crossroads. *Journal of Spatial and Organisational Dynamics*, *4*(3), 199–212.

Guidry, J. P. D., Austin, L. L., Carlyle, K. E., Freberg, K., Cacciatore, M., Meganck, S., Jin, Y., & Messner, M. (2018). Welcome or Not: Comparing #Refugee Posts on Instagram and Pinterest. *American Behavioral Scientist*, *62*(4), 512–531. https://doi.org/10.1177/0002764218760369

Guynn, J. (2019). Facebook while black: Users call it getting "Zucked", say talking about racism is consored as hate speech. *USA Today News*. https://eu.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002/

GWI. (2022). *Social. GWI's flagship report on the latest trends in social media*. https://www.gwi.com/reports/social

Hasell, A. (2021). Shared Emotion: The Social Amplification of Partisan News on Twitter. *Digital Journalism*, *9*(8), 1085–1102. https://doi.org/10.1080/21670811.2020.1831937

Heidenreich, T., Eberl, J.-M., Lind, F., & Boomgaarden, H. (2020). Political migration discourses on social media: A comparative perspective on visibility and sentiment across political Facebook accounts in Europe. *Journal of Ethnic and Migration Studies*, *46*(7), 1261–1280. https://doi.org/10.1080/1369183X.2019.1665990

Heilweil, R. (2020). Facebook is flagging some coronavirus news posts as spam. *Vox*. https://www.vox.com/recode/2020/3/17/21183557/coronavirus-youtube-facebook-twitter-social-media

Hern, A. (2020). *Twitter apologises for racist" image cropping algorithm*. https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

Hsueh, M., Yogeeswaran, K., & Malinen, S. (2015). "Leave Your Comment Below": Can Biased Online Comments Influence Our Own Prejudicial Attitudes and Behaviors?: Online Comments on Prejudice Expression. *Human Communication Research*, *41*(4), 557–576. https://doi.org/10.1111/hcre.12059

Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2021). Algorithmic amplification of politics on Twitter. *Proceedings of the National Academy of Sciences*, *119*(1), e2025334119. https://doi.org/10.1073/pnas.2025334119

Impronta, R. (2020). *Instagram nuova policy censura contenuti espliciti: Influencer indignati*. https://informa-press.it/instagram-policy-censura-influencer/

Jaramillo-Dent, D., Contreras-Pulido, P., & Pérez-Rodríguez, M. A. (2022). Right-wing immigration narratives in Spain: A study of persuasion on Instagram Stories. *European Journal of Communication*, *37*(2), 161–180. https://doi.org/10.1177/02673231211012157

Jaso, M. (2022). How War in Ukraine Roiled Facebook and Instagram. *The New York Times*. https://www.nytimes.com/2022/03/30/technology/ukraine-russia-facebook-instagram.html

Joles, B. (2022). *Inside the risky world of "Migrant TikTok."* Rest of World. https://restofworld.org/2022/migrant-tiktok-flourishes/?mc_cid=8c3921687f&mc_eid=e50688ab33

Joslin, T. (2020). *Black creators protest TikTok's algorithm with #ImBlackMovement*. https://www.dailydot.com/irl/tiktok-protest-imblackmovement/

Kaufmann, K. (2018). Navigating a new life: Syrian refugees and their smartphones in Vienna. *Information, Communication & Society*, *21*(6), 882–898. https://doi.org/10.1080/1369118X.2018.1437205

Kayser-Bril, N. (2020). *Automated discrimination: Facebook uses gross stereotypes to optimize ad delivery*. https://algorithmwatch.org/en/automated-discrimination-facebook-google/

Kelly, M. (2020). *TikTok pledges to promote black creators after accusations of censorship*. https://www.theverge.com/2020/6/1/21277505/tiktok-black-creators-censorship-algorithm-donation-diversity-council

Kilpatrick, J., & Jones, C. (2022). *A clear and present danger Missing safeguards on migration and asylum in the EU's AI Act*. Statewatch. https://www.statewatch.org/media/3285/sw-a-clear-and-present-danger-ai-act-migration-11-5-22.pdf

Klein, O. (2020). How is the far right capitalizing on COVID-19? *Centre for Analysis of the Radical Rights*. https://www.radicalrightanalysis.com/2020/04/10/how-is-the-far-right-capitalizing-covid-19/

Knight, W. (2021). *Twitter's Photo-Cropping Algorithm favors young, thin females*. https://www.wired.com/story/twitters-photo-cropping-algorithm-favors-young-thin-females/

Kreis, R. (2017). #refugeesnotwelcome: Anti-refugee discourse on Twitter. *Discourse and Communication*. https://doi.org/10.1177/1750481317714121

Kuo, L. (2019). *TikTok sorry for blocking teenager who disguised Xinjiang video as make-up tutorial*. https://www.theguardian.com/technology/2019/nov/28/tiktok-says-sorry-to-us-teenager-blocked-after-sharing-xinjiang-videos

Kwon, K. H., & Gruzd, A. (2017). Is offensive commenting contagious online? Examining public vs interpersonal swearing in response to Donald Trump's YouTube campaign videos. *Internet Research*, *27*(4), 991–1010. https://doi.org/10.1108/IntR-02-2017-0072

Ladner, K., Ramineni, R., & George, K. M. (2019). Activeness of Syrian refugee crisis: An analysis of tweets. *Social Network Analysis and Mining*, *9*(1), 61. https://doi.org/10.1007/s13278-019-0606-6

Lambrecht, A., & Tucker, C. E. (2018). Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2852260

Latorre, J. P., & Amores, J. J. (2021). Topic modelling of racist and xenophobic YouTube comments. Analyzing hate speech against migrants and refugees spread through YouTube in Spanish. *Ninth International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM'21)*, 456–460. https://doi.org/10.1145/3486011.3486494

Lee, J. J., & Mccabe, J. M. (2021). Who Speaks and Who Listens: Revisiting the Chilly Climate in College Classrooms. *Gender & Society*, *35*(1), 32–60. https://doi.org/10.1177/0891243220977141

Lee, J.-S., & Nerghes, A. (2018). Refugee or Migrant Crisis? Labels, Perceived Agency, and Sentiment Polarity in Online Discussions. *Social Media + Society*, *4*(3), 205630511878563. https://doi.org/10.1177/2056305118785638

Leung, L. (2010). Telecommunications across borders: Refugees' technology use during displacement. *Telecommunications Journal of Australia*, *60*(4), 51–58.

Leurs, K., & Smets, K. (2018). Five Questions for Digital Migration Studies: Learning From Digital Connectivity and Forced Migration In(to) Europe. *Social Media + Society*, *4*(1), 205630511876442. https://doi.org/10.1177/2056305118764425

MacCarthy, M. (2021). *How online platform transparency can improve content modearation and algorithmic performance*.

McCluskey, M. (2020). These TikTok Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content. *Time*. https://time.com/5863350/tiktok-black-creators/

Meaker, M. (2018). Europe is using smartphone data as a weapon to deport refugees. *Wired*. https://www.wired.co.uk/article/europe-immigration-refugees-smartphone-metadata-deportations

Meta. (2018a). *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*. https://about.fb.com/news/2018/11/myanmar-hria/

Meta. (2018b). *Update on Myanmar*. https://about.fb.com/news/2018/08/update-on-myanmar/

Meta. (2021). *How Does News Feed Work? Episode 3 of Let me Explain Has Answers.* https://www.facebook.com/business/news/let-me-explain-video-series-how-does-news-feed-work

Milano, S., Taddeo, M., & Floridi, L. (2020). Recommender systems and their ethical challenges. *AI & SOCIETY*, *35*(4), 957–967. https://doi.org/10.1007/s00146-020-00950-y

Mozilla. (2021). *YouTube Regrets. A crowdsourced investigation into YouTube's recommendation algorithm*. https://foundation.mozilla.org/en/youtube/findings/

Murphy, L., & Cacace, M. (2020). *Facebook's Civil Rights Audit – Final Report*. Facebook.

Murray, K. E., & Marx, D. M. (2013). Attitudes toward unauthorized immigrants, authorized immigrants, and refugees. *Cultural Diversity and Ethnic Minority Psychology*, *19*(3), 332–341. https://doi.org/10.1037/a0030812

Nerghes, A., & Lee, J.-S. (2018). The Refugee/Migrant Crisis Dichotomy on Twitter: A Network and Sentiment Perspective. *Proceedings of the 10th ACM Conference on Web Science*, 271–280. https://doi.org/10.1145/3201064.3201087

Nerghes, A., & Lee, J.-S. (2019). Narratives of the Refugee Crisis: A comparative Study of Mainstream-Media and Twitter. *Media and Communication*, *7*(20). https://www.cogitatiopress.com/mediaandcommunication/article/view/1983%253B/0

Poell, T., Nieborg, D., & Brooke Erin, D. (2022). *Platforms and Cultural Production*. Polity Press.

Privacy International. (2017). *Social Media Intelligence*. https://privacyinternational.org/explainer/55/social-media-intelligence

Rankin, J., & MacDowell, R. (2021). How to overcome Zoom's algorithmic bias. *Teaching + Learning Lab*. https://tll.mit.edu/how-to-overcome-zooms-algorithmic-bias/

Reuters. (2022). Facebook and Instagram let users call for death to Russian soldiers over Ukraine. *The Guardian*. https://www.theguardian.com/technology/2022/mar/11/facebook-and-instagram-let-users-call-for-death-to-russian-soldiers-over-ukraine

Roberts, Z. (2017). Information Exchange between Smugglers and Migrants: An Analysis of Online Interactions in Facebook Groups. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3051186

Rosina, M. (2022). *The Criminalisation of Irregular Migration in Europe. Globalisation, Deterrence and Vicious Cycles*. Palgrave Macmillan.

Ryan, F., Fritz, A., & Impiombato, D. (2020). *TikTok and WeChat Curating and controlling global information flows*. ASPI. https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2020-09/TikTok%20and%20WeChat.pdf?7BNJWaoHImPVE_6KKcBP1JRD5fRnAVTZ=

Sap, M., Card, D., Gabriel, S., Choi, Y., & A. Smith, N. (2019). *The Risk of Racial Bias in Hate Speech Detection*. ACL. https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf

Sawyer, R., & Chen, G. (2012). The impact of social media on intercultural adaptation. *Intercultural Communication Studies*, *21*, 151–169.

Schafer, C., & Schadauer, A. (2018). *Online Fake News, Hateful Posts Against Refugees, and a Surge in Xenophobia and Hate Crimes in Austria*. Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781351049634-13/online-fake-news-hateful-posts-refugees-surge-xenophobia-hate-crimes-austria-claudia-sch%C3%A4fer-andreas-schadauer

Schemer, C., & Meltzer, C. E. (2020). The Impact of Negative Parasocial and Vicarious Contact with Refugees in the Media on Attitudes toward Refugees. *Mass Communication and Society*, *23*(2), 230–248. https://doi.org/10.1080/15205436.2019.1692037

Schmidle, N. (2015, October 19). Ten Borders: One refugee's epic escape from Syria. *The New Yorker: One Refugee's Epic Escape from Syria*. https://www.newyorker.com/magazine/2015/10/26/ten-borders

Schroepfer, M. (2019). *Community Standards report*. https://ai.facebook.com/blog/community-standards-report/

Siapera, E., Boudourides, M., Lenis, S., & Suiter, J. (2018). Refugees and Network Publics on Twitter: Networked Framing, Affect, and Capture. *Social Media + Society*, *4*(1), 205630511876443. https://doi.org/10.1177/2056305118764437

Silverman, C. (2020). Black Lives Matter Activists Say They're Being Silenced By Facebook. *BuzzFeed News*. https://www.buzzfeednews.com/article/craigsilverman/facebook-silencing-black-lives-matter-activists

Solomon, R. S., Pykl, S., Das, A., Gamback, B., & Chakraborty, T. (2019). Understanding the Psycho-Sociological Facets of Homophily in Social Network Communities. *IEEE Computational Intelligence Magazine*, *14*(2), 28–40. https://doi.org/10.1109/MCI.2019.2901084

Solon, O. (2020). "Facebook doesn't care": Activists say accounts removed despite Zuckerberg's free-speech stance. *NBC News*. https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110

Spörlein, C., & Schlueter, E. (2021). Ethnic Insults in YouTube Comments: Social Contagion and Selection Effects During the German "Refugee Crisis." *European Sociological Review*, *37*(3), 411–428. https://doi.org/10.1093/esr/jcaa053

Stack, L. (2018). What is a "shadowban" and is Twitter Doing it to republican accounts? *The New York Times*. https://www.nytimes.com/2018/07/26/us/politics/twitter-shadowbanning.html

Statista. (2022a). *Daily time spent on social networking by internet users worldwide from 2012 to 2022*. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Statista. (2022b). *Most popular social networks worldwide as of January 2022, ranked by number of monthly active users*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Tidey, A. (2018). High Facebook usage linked to anti-refugee attacks in Germany: Study. *Euronews*. https://www.euronews.com/2018/08/22/high-facebook-usage-linked-to-anti-refugee-attacks-in-germany-study

Tobin, A., & Merrill, J. B. (2018). *Facebook Is Letting Job Advertisers Target Only Men*. https://www.propublica.org/article/facebook-is-letting-job-advertisers-target-only-men

Tracking.Exposed project. (2022). *Mapping Ban and Shadow-Ban on TikTok: Expose hidden censorship with a cross-national research*. https://tiktok.tracking.exposed/ws22-shadowban-research/

Twitter. (2022). *About your Twitter timeline. What's in yout Home timeline*. https://help.twitter.com/pl/using-twitter/twitter-timeline

UN. (2020). *Global issues: Migration*. https://www.un.org/en/global-issues/migration

UN Migration. (2019). *Responding to hate speech against migrants on social media: What can you do?* https://rosanjose.iom.int/en/blogs/responding-hate-speech-against-migrants-social-media-what-can-you-do

UNHCR. (2016). *Connecting Refugee. How Internet and Mobile Connectivity can Improve Refugee Well-Being and Transform Humanitarian Action*. https://www.unhcr.org/5770d43c4.pdf

Urchs, S., Wendlinger, J., Flynn, J. R., & Granitzer, M. (2019). A Twitter Dataset for Extracting and Analysing Migration-Movement Data of the European Migration Crisis 2015. *IEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterpreises (WETICE)*, 146–149. https://doi.org/10.1109/WETICE.2019.00039.

Urman, A., Makhortykh, M., & Ulloa, R. (2022). Auditing the representation of migrants in image web search results. *Humanities and Social Sciences Communications*, *9*(1), 130. https://doi.org/10.1057/s41599-022-01144-1

van Klingeren, M., Boomgaarden, H. G., Vliegenthart, R., & de Vreese, C. H. (2015). Real World is Not Enough: The Media as an Additional Source of Negative Attitudes Toward Immigration, Comparing Denmark and the Netherlands. *European Sociological Review*, *31*(3), 268–283. https://doi.org/10.1093/esr/jcu089

Vincent, J. (2020). Facebook is now using AI to sort content for quicker moderation. *The Verge*. https://www.theverge.com/2020/11/13/21562596/facebook-ai-moderation

Walker, S. (2021). *What is The Difference Between Automated and Human COntent Moderation*. https://newmediaservices.com.au/automated-and-live-moderation/

Wall, M., Campbell, M. O., & Janbek, D. (2017). Syrian refugees and information precarity. *New Media & Society*, *19*(2), 240–254.

Wright, C., Brinklow-Vaughn, R., Johannes, K., & Rodriguez, F. (2021). Media Portrayals of Immigration and Refugees in Hard and Fake News and Their Impact on Consumer Attitudes. *Howard Journal of Communications*, *32*(4), 331–351. https://doi.org/10.1080/10646175.2020.1810180

Yantaseva, V. (2020). Migration Discourse in Sweden: Frames and Sentiments in Mainstream and Social Media. *Social Media + Society. October*. https://doi.org/doi:10.1177/2056305120981059

Zakrezewski, C. (2020). The Technology 202: Instagram faces backlash for removing posts supporting Soleimani. *The Washington Post*. https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/01/13/the-technology-202-instagram-faces-backlash-for-removing-posts-praising-soleimani/5e1b7f1788e0fa2262dcbc72/