Annemarie Witschas

Porn, Power, and Platforms

The (re-)production of hegemonic sexuality through Machine Learning systems

PICS 2023, Number 1

PICS
Publications of the
Institute of Cognitive Science

PICS
Publications of the Institute of Cognitive Science

Annemarie Witschas. 2023. *Porn, Power, and Platforms. The (re-)production of hegemonic sexuality through Machine Learning systems* (Publications of the Institute of Cognitive Science 2023, Number 1). Osnabrück: Institute of Cognitive Science, Osnabrück University.

This title can be downloaded at:
https://osnadocs.ub.uni-osnabrueck.de

Institut für Kognitionswissenschaft
Universität Osnabrück
49069 Osnabrück
Germany
https://ikw.uni-osnabrueck.de

Storage and cataloging done by Osnabrück University

Foreword to "Porn, Power, and Platforms. The (re-)production of hegemonic sexuality through Machine Learning systems" (Annemarie Witschas)

Artificial Intelligence not only permeates almost all areas of our daily lives, but also academic discourse, even beyond the computer and engineering faculties. What is needed to shape this development responsibly are ethical assessments of AI technology that go beyond the standard topics of self-driving cars and robots in elderly care or education. Focusing ethical attention instead on AI applications that are less public and visible, such as those in the field of pornography, is as considerable a challenge as it is necessary. In her bachelor thesis, Annemarie Witschas demonstrates that cognitive science, when done in a truly interdisciplinary spirit, can contribute substantially to ethics and critique of new technology. What makes her approach especially praiseworthy is her competence to explain and analyze technological systems, while at the same time drawing on the repertoire of critical and social philosophy to examine the social and normative implications of these artifacts.

The innovative approach of Witschas' bachelor thesis consists in her exploration of the entanglement between AI technology and social power. Witschas shows how the subject of online pornography provides a lens through which to analyze the complex relationship between AI technology and social norms alongside power relations. Especially within the interdisciplinary context of cognitive science does Witschas' thesis present an impressive scholarly achievement, as it credibly transcends disciplinary horizons in its inclusion of a remarkable corpus of literature ranging from Queer Studies, Social Philosophy, and Machine Learning to Porn Studies. In this way, the achievement is not only impressive, but exemplary in the field of cognitive science.

With its decidedly "critical stance" that aims at the flourishing and freedom of humans to express themselves in a self-determined and diverse manner, Witschas' thesis is highly empowering in its mode of discourse. It cogently analyzes the avenues through which recent technologies can implicitly act as restrictive and oppressive apparatuses in realms such as that of sexuality, showing that these obstacles to diversity and self-determination are *made* and thus could also be re-made. We take Witschas' thesis to be an inspiring work for future cognitive science students who might in a similar spirit embark on critical and empowering analyses of urgent societal matters in their relation to our own discipline.

Osnabrück, March 2023

Rainer Mühlhoff Imke von Maur

Cognitive Science B.Sc.

Porn, Power, and Platforms

The (re-)production of hegemonic sexuality through Machine Learning systems

Revised version of the Bachelor Thesis submitted by

Annemarie Witschas
awitschas@uos.de

Supervised by

Dr. Rainer Mühlhoff
Cluster Science of Intelligence, Technische Universität Berlin

Dr. Imke von Maur
Institut für Kognitionswissenschaft, Universität Osnabrück

Contents

1 Introduction

Pornography is a polarizing topic. Pleasure and entertainment to some, while simply distasteful or even dangerous to others, who postulate its malignant potential of causing addiction, spoiling adolescents' minds and disrupting marriages (The Recovery Village, 2021). Likewise, feminist discourses are quite divided on the matters of pornography. For some, it represents the epitome of patriarchy, enacting male domination and female subordination, promoting a male gaze, neglecting topics of female pleasure and generating billions of dollars of profit for white men (e.g. MacKinnon, 1989; Dworkin, 1985; Dines et al., 1998). Others see not pornography itself as evil but the practices shaping it. They identify the creation of feminist pornographic material not only as source of pleasure but also of potential liberation (Liberman, 2015; Shrage, 2005; Taormino et al., 2013).

In spite of the controversies surrounding it, pornography is a topic too large and significant to be left aside. And with today's massive influence of the internet on everyday life the significance of online pornography becomes even more paramount. Already in the early 2000s, the musical 'Avenue Q' had chantingly proclaimed on Broadway stages that 'the internet is for porn' (Bähr, 13.02.2015). Indeed, online pornography is not just an often recited meme but in fact an important phenomenon of today's digital cultures. The substantial role of pornography on the internet becomes palpable in statistics of online traffic: four of the 20 most visited websites are pornography platforms (SimilarWeb, 22.03.2021), each one receiving around 3 billion visits per month (Neufeld, 27.01.2021). PornHub, at time of writing the third most visited porn site welcomes more than a 100 million visitors each day (Pornhub Insights, 2019; SimilarWeb, 22.03.2021). Reliable and up-to-date empirical evidence for the frequency of *online* pornography consumption is hard to find. For orientation, a US study from 2018 found that 91.5 % of men and 60.2 % of women had consumed online or offline pornography in the past month (Solano et al., 2020).[1] Still, porn and particularly online porn is not adequately represented in public discourses and scientific theories (Paasonen, 2011). There is, one might assume, a veil of indecency still lingering over pornography caused by a stigma on sexuality.

This work does not try to argue for or against the consumption of pornography, it does not go further than acknowledging the heterogeneity of practices that can have varying liberatory or repressive effects. Rather, the phenomenon of pornography is merely *used as a lens* through which I attempt to portray the entanglement of Artificial Intelligence (AI) and power.

AI and its subtype Machine Learning (ML) are widely applied technologies that are entering more and more aspects of life, including health care, credit rating, policing, jurisdiction or surveillance (O'Neil, 2016; Mehrabi et al., 2019). With the dissemination of AI in social and political spheres questions about the dimension of power in AI are becoming

[1] N = 1392, ages 18-73, including written pornography, pictures and videos.

more pressing. In the past years, researchers have uncovered gendered and racialized biases in various types of AI models (e.g. Buolamwini and Gebru, 2018; Mehrabi et al., 2019; Leavy, 2018). Besides perpetuating inequality through bias, ML systems also pose another threat: they are involved in the (re-)production of certain norms, subjects and cultures. These norms could for instance be severely retrogressive, aligned with 'traditional' gender roles, heteronormative and cis-sexist conceptions.

A starting point for this work might be a reflection on the paradoxical attitude toward pornography in our digital cultures: On the one side, we are frenetically eradicating any kind of visual evidence of human sexuality from social media. At the same time, tube-site platforms that contain toxic practices, unconsensually uploaded material and precarious labor conditions have become the established go-to place for persons seeking sexual pleasure online. *Why is that the case?* Why are consensual, heterogeneous and creative sexual practices so rare online? How can it be that the internet—a space that once promised open possibilities and free participation—is now dominated by a small set of potentially retrogressive values? What role does the application of AI, or more precisely ML, play in this matter?

Centrally, the question that I examine in this work is how power operates through ML systems in order to manifest hegemonic sexual norms. For this investigation, I consider two artefacts at the intersection of ML and pornography. The first one, content moderation, is primarily centred on removing undesired content such as pornography from social media platforms. Thus, it does not merely operate negatively, by removing what counts as pornography, but it also produces and reproduces a working definition of the very concept it is trying to grasp. Deepfakes, the second application, are manipulated videos in which arbitrary faces are inserted through the use of Deep Learning (DL) to create realistic footage. Deepfakes have recently received a lot of public attention, mostly for the epistemic threat they pose to democracies (e.g. Fallis, 2020). However, the number one application for deepfakes is the generation of non-consensual pornography (Ajder et al., 2019).

I situate this work within a context of critical AI studies and use concepts and insights from different thinkers and theoretical frameworks. First of all, I apply a socio-technical understanding of ML technologies which denies that technology is merely a neutral tool but holds and (re-)produces values and norms. To understand how ML systems can have the ability to shape social spheres, a notion of power is needed: the conception of power that I apply is inspired by the work of the philosopher Michel Foucault who emphasizes a relational, distributed and productive mode of power. Central to this work are also the insights of queer theory and its critique of the normativities shaping sexualities and gender identities. Through this queer perspective, particularly matters of normativity and deviance come into focus and blend with Foucault's concept of productive power. By incorporating Sara Ahmed's notion of straightening devices I examine how the use of ML applications form and create sexual subjects, desires and cultures. In addition, this

work also draws upon the concept of cultural hegemony developed by Marxist philosopher Antonio Gramsci that makes intelligible how power is sustained through the 'consent of the masses' (Gramsci, 2011, p. 145).

This work is subdivided into four major parts. The first part contains a clarification and delineation of the most important concepts. Secondly, I introduce two ML systems and their application for the generation and deletion of online pornography. This part also includes a brief explanation of the technical mechanisms that these systems are based on. The third part analyzes how power is exerted through them: After recapitulating the conception of power based on the work of Michel Foucault and Sara Ahmed, I investigate the ability of these systems to promote "traditional" (hegemonic) sexual norms, considering concrete effects of these ML systems. Finally, I shortly consider the influential positions of platforms as the sites where multiple elements of ML systems intersect. As I observe, large commercial platforms offer optimal conditions for the propagation of hegemonic norms.

2 Concepts

To begin, I clarify and delineate the most important concepts used throughout this work.

2.1 Machine/Deep Learning

Artificial Intelligence, quite generally, aims at developing computational programs that humans deem as intelligent, be it playing chess, recognizing objects or having conversations. Early approaches to AI attempted to solve this by creating precise instructions for computers to follow. In order to accurately formalize the problem at hand developers had to explicitly specify rules for a respective task. While those so-called "symbolic approaches" (Dick, 2019) were strikingly successful with problems like chess, they turned out to be rather untenable for other—more banal—tasks. In particular, tasks which humans can solve quite easily can be extremely hard for them to describe formally (Goodfellow et al., 2016, p. 1). The subfield of Machine Learning (ML), thus, aimed at having the computational model infer underlying principles by itself. To do so, ML usually requires vast amounts of data to detect correlations or other regularities among the data. Hence, these approaches can be considered forms of *data-driven* AI. Fueled by the sharp increase of available data over the last decades, ML has advanced to an extensively applied technology. The combination of rapidly developing ML technology, affordable computational power and availability of massive amounts of data denote the character of *contemporary* AI. Deep Learning (DL), then, is a method of ML that uses a specific kind of architecture: artificial neural networks. Inspired by the operation of neurons in the brain, these networks are formalized by often millions of parameters, commonly ordered in layered structures of smaller computational units, referred to as 'neurons'. Through this complex model architecture, DL 'achieves great power and flexibility by representing the world as

a nested hierarchy of concepts, with each concept defined in relation to simpler concepts' (Goodfellow et al., 2016, p. 8).

Instead of focussing on ML applications merely as technological artefacts, I consider them as **socio-technical systems**. This perspective allows to more closely examine how technology like ML is intertwined with its broader societal context. Firstly, technologies like ML applications are always brought about by a specific societal context and therefore are reflecting its present values and norms. A given technological artefact can thus tell much about the society that conceived it. The specific framing of a problem and the methodology it uses to solve this problem are indicative of the dominant conceptions and prevalent knowledge system (*'episteme'*), which are in themselves closely tied to power (Foucault, 2012, p. 27; 2005, p. 138). Hence, technologies can never be conceived as neutral, but inevitably inherit values of their developers, commissioners and users (Cobbe, 2020). Secondly, deployed technologies are also non-neutral in regard to their effects on society. They shape society and its conceptions and are implicated in the production of certain discourses, material conditions and subjects, as I explore in the sections 4.4.1 and 4.4.2 for the examples of content moderation and deepfakes.

Consequently, the exact operation of ML applications and the effects they bring about strongly depend on their interaction with other elements. To analyse content moderation and deepfakes as ML systems therefore requires including the following elements: Firstly, ML systems employ an algorithm or, precisely speaking, a model. In its technical definition, an *algorithm* can be understood as an executable and finite definition of an order of operations (e.g. Rapaport, 2012, p. 72). Algorithms are not proprietary to computers but can be performed by humans, too. For instance, cooking recipes or manuals are also algorithms by definition. Although often used synonymously, *models* differ from algorithms. When the procedure of the algorithm is executed and instantiated with real values, it creates a model. A model, then, is the result of applying the operations specified by an algorithm onto concrete data (Brownlee, 2016, p. 10). *Data* is the next crucial component for ML systems and one where the interconnection of social spheres with technical functionality is particularly conspicuous. As data-driven forms of AI, the performance of ML and DL models crucially depends on large data sets that constitute the interface between social spheres and the digital infrastructure. However, data inherently embodies a selective representation of the world and can thus amplify existing biases or induce novel ones (Mehrabi et al., 2019). Hence, the results of ML applications depend on how data was gathered, and by whom. In many cases, humans contribute with paid or unpaid labor to the creation of data sets used for ML (Mühlhoff, 2021) and imprint it with their biases and cultural contingencies. Further, in both content moderation and deepfakes, *platforms* play a central role. I use the term platforms to refer to online sites operated by commercial agents that enable networking and sharing of 'user-generated content' (Roberts, 2018) among users. The platforms I focus on are social networking and social media sites,

like Tumblr, Instagram, YouTube, and PornHub[2]. Platforms are fundamentally tied to content moderation and also provide spaces to develop, improve, disseminate and monetize pornographic deepfakes. In their primary role, platforms are intermediaries between different parties, including users, advertisers and other platforms. Yet, platform administrations are not only passive hosts of user-generated content, but also actively shape what can be experienced online through the filtering and ranking of content. As the ones who create and upload content on platforms, *users* distinctly shape the mode of operation of platforms and create demand for moderation. Economic incentives and particularly the appeal to *advertisers* also shape the policies of platforms and can have far-reaching consequences on marginalized communities, as I am to stress in section 4.4. Moreover, ML systems like content moderation and pornographic deepfakes are affected by the existence of *legislations*, or lack thereof, which can either enhance or restrict their functioning.

This list is by no means comprehensive, countless more elements that shape the functioning of the ML systems at hand could be identified and included. In regard to their relevance and the scope of this work, I focus on the ones above.

2.2 Hegemonic sexuality

The second component in my argumentation will be 'hegemonic sexuality'. The term 'hegemony' dates back to Marxist philosopher Antonio Gramsci and refers to a position of power and dominance within a society. I focus on the part of *cultural* hegemony which Gramsci describes as '[t]he "spontaneous" consent given by the great masses of the population to the general direction imposed on social life by the dominant fundamental group' (Gramsci, 2011, p. 145).[3] Gramsci points to the existence of certain conceptions that are pre-reflectively accepted by the majority of people. These conceptions are perceived as the "default" or "normal way of doing things" and are thus seldomly questioned. However, these conceptions might be historically and socially contingent, established at a certain point in time and benefiting specific groups while disadvantaging others.

Generally, norms prescribe whether certain actions are demanded, allowed or forbidden (von Kutschera, 1973, p. 11), transgressing them is socially penalized. Hegemonic norms, then, are the dominant conceptions present in a society that find general approval in the population and which serve to uphold existing asymmetries of power. Since my focus in this work will be on pornography, I specifically limit my analysis to hegemonic norms in the realm of sexuality. Yet, I understand the phenomenon of sexuality more broadly, including also gendered norms and the material conditions shaping the practice of porn production. I summarize these norms under the term 'hegemonic sexuality'. This term is inspired by and certainly partly overlapping with the term of 'hegemonic masculinity',

[2]Although the type of content shared on PornHub is thematically more narrow, it operates under the same principle of hosting content uploaded by users as the other platforms in the list.

[3]To fit the scope of this work, I set aside questions about the 'dominant fundamental group' and leave out an analysis of class relations, although it certainly is insightful to the greater phenomenon.

as coined by Connell (1998). Whereas hegemonic masculinity in essence refers to *male gendered* norms that legitimize and sustain a male dominance over other genders (e.g. Donaldson, 1993), hegemonic sexuality encapsulates *all genders* into its analysis. Thus, I conceive hegemonic sexuality as encompassing norms of sexual expressions, practices and identities which aim to reproduce existing relations of forces.[4] Examples for hegemonic sexuality would be:

1. Heteronormativity, the perception of heterosexuality as normal, the default, while perceiving other sexualities as deviant.

2. Normativity of monogamy, the idea that sexuality should take place in closed relationships confined to exactly two people.

3. A binary-sexed and able-bodied body norm.

4. Gendered sexual roles, i.e., the conception of men having stronger sexual drives and thus acting as "hunters" whereas female sexuality is constructed as a precious object to be concealed and protected from the force of male sexuality. Hence, 'slutshaming' or belittling men for their sexual inexperience are forms of the hegemonic sexuality.

The argumentation of these examples as historically contingent, but nonetheless widely accepted norms has been done elsewhere (see e.g. Ahmed, 2006; Pieper and Bauer, 2014; hooks, 01.07.1999) and will not be further elaborated here.

In this work, I avoid rigid and binary gender terminology like "women" and "men". Instead, I use the term FLINTA* (Female Lesbian Intersex Non-Binary Transgender Agender *) to include a variety of non-cis-male genders into my analysis. This term specifically refers to the *gender identity* of a person. However, when talking about deepfakes, the gender identity of the person depicted is often unknown. To refrain from concluding someone's gender from their facial features, I instead use the term *femme-presenting* to refer to persons who carry markers associated with female gender and who are socially "read" as women, although they might identify differently (see also Wagner and Blewer, 2019). As a note of caution, this terminology may implicitly reproduce a binary gender understanding for it does not challenge the existent practice of sorting people into two categories based on their appearance. However, to make the systemic patterns of harassment in deepfake porn visible it is nonetheless important to use a term that denotes gender *appearance* additionally to identity.

[4]Rooted in this term is an understanding of gender and sexuality as social constructs founded contingently upon capitalist accumulation strategies (e.g. Federici, 2004), which I however will not further dive into.

3 Machine Learning applications for porn

In the following section, I introduce two ML systems, content moderation and deepfakes. Both of them are not solely applied to pornography, but are tools that can be utilized for various purposes. For instance, content moderation finds application in handling hate speech or copyright infringements, too. Deepfakes are also used for comedic purposes and, to a minor degree, for manipulating videos of politicians Ajder et al. (2019). Nonetheless, pornography is a central use case to both of them. Hence, both fundamentally contribute to shaping the online landscape of pornography as well as a societal understanding of sexuality.

3.1 Content moderation

Content moderation refers to the process of reviewing content posted on online platforms with the aim of filtering out illicit content (Singh, 2019; Gillespie, 2018). Reasons for the removal of online content can be copyright infringements, graphic violence, hate speech or pornography (Singh, 2019).

3.1.1 Types of content moderation

There are three main pillars involved in realizing the process of content moderation: hired staff, voluntary user efforts and automation technologies. These can be combined to varying degrees by different platforms (Gillespie, 2018, pp. 78–110). Today, a significant share of content moderation is outsourced to contracted 'click workers', commonly located in the Global South due to low wages (Chen, 23.10.2014). Yet, the enormous amounts of content posted online daily and the high cost of human labor make a purely manual revision process untenable for most platforms. For illustration, every minute 243,000 photos are uploaded on Facebook (Omnicore, 2021) and 2.8 hours of video are added to PornHub (Pornhub Insights, 2019). In order to keep up with the amount of uploads, most platforms automate the process of content moderation to some degree (e.g. Pornhub, 03.04.2021; Tumblr Support, 2018).

Moderation policies, or the rules that state which content is permitted are usually specified by the platform's administration and pertain all over the platform. Yet, these *global* policies can also be supplemented with *local* policies (Singh, 2019). On Reddit, for instance, user moderators are commissioned with the oversight of a specific subchannel (subreddit) and can also impose new rules on this part of the site (Singh, 2019). Creators of a Facebook page for a business or organization can also specify a list of certain words to be banned on their page (Facebook for Media, 2021).

Content moderation is not restricted to solely performing binary decisions. Instead of removing an unwanted post, moderation systems can also influence its ranking and visibility on the platform. This practice recently sparked controversy when leaked documents

from TikTok revealed that moderators were instructed to lower the visibility of videos that featured disabled, queer or obese people (Köver, 02.12.2019).[5]

3.1.2 Difficulties of content moderation

Notably, the decision making process in content moderation is often subjective, contextual and hard to formalize for automated models (Gillespie, 2018; Roberts, 2018). Without the ability to properly evaluate nuances and contextual information, automated approaches have a high error rate, failing to discern irony, newsworthy content or art from profanities. The distinction of these is non-trivial even for humans and contingent based on cultural contexts. For instance, certain pejorative words might be used to exert violence over an oppressed group and should—in alignment with content policies—be taken off the platform. However, these words can also be reclaimed by an oppressed group, their reinterpretation acting as an empowerment over hostility (Şahin, 2019). For instance, 'slut' is a word historically used as a slur, penalizing females for an active or outspoken sexuality that is inconsistent with the hegemonic sexuality intended for them. Today, this word has been reclaimed and assigned a new meaning by activists, denoting self-determined sexual subjects of any gender (Easton and Liszt, 1997). Finding ambiguous words like 'slut' in posts online, content moderation models have no way in interpreting the emancipating or repressive intentions of the speech act, even less its material consequences, and thus fail to properly classify it as legitimate or illegitimate. Similarly, Dias Oliva et al. (2021) reported that 'Perspective', an AI technology developed by Google to detect toxic language, rated the toxicity of famous drag queens higher than those of far-right conservatives, as the software was unable to account for 'mock impoliteness', a strategy of LGBTQ* people to cope with hostility.

Another illustrative example of the importance of contextual information in content moderation is the Pulitzer price winning picture of Phan Thi Kim Phuc, also known as the Napalm girl. Facebook was harshly criticized when the famous photo depicting the naked girl running away from a bomb attack during the Vietnam war was repeatedly removed from its platform (Gillespie, 2018; Roberts, 2018). This photo sparked a debate on how to handle content that unquestionably infringes moderation policies, e.g. by depicting cruelty and nudity, without eradicating important historical artefacts.[6] The correct evaluation of context and cultural significance of an uploaded image is a challenge for both automated approaches as well as hired moderators, who only have a few seconds to judge an image in front of them (Gillespie, 2018, p. 112).

[5] Videos depicting these bodies would not be allowed into the 'For-You-Feed', the algorithmically compiled start page where TikTok users can reach the largest audiences. Further, their visibility was confined to their home country and could not be seen worldwide as other TikTok videos. While TikTok has argued that this policy is aimed to protect 'vulnerable' users from bullying, other documents also indicate that the platform reduces their visibility in order to enhance their appeal to new users and advertisers (Biddle et al., 16.03.2020).

[6] Central to this debate are also the matters of cultural imperialism that influence which cultural backgrounds are regarded valuable.

Apart from the undoubted complexity of some decisions, automated content moderation systems also suffer from being incredibly easy to fool with banal distractions. For instance, Tumblr users noticed that a previously flagged image would no longer be recognized as "adult" content by the moderation model after adding a small distractor picture at its bottom (Ellison, 07.12.2018).

3.1.3 Automated approaches of pornography moderation

How can computers know if an image contains pornography? And how do major social media platforms implement pornography detection? The second question is difficult to answer, given that concrete algorithms are scarcely ever made publicly available, often as a way of protecting the platforms' "business secrets". Thus, I can only cover general aspects of the procedure that should nonetheless illustrate on which grounds automated pornography detection operates and what are reasons for its failures.

I focus on image- and video-based content as these are the types of pornography that are most commonly censored (e.g. Tumblr Support, 2018). One computationally rather simple approach is to assign a hash function that can be conceived as the 'digital fingerprint' of a picture (Llansó et al., 2020). This method called 'hashing' is primarily used to compare newly uploaded images to existing databases of harmful content. One of such databases stores child exploitation material and can thus prevent the re-upload of these images (Llansó et al., 2020). To detect previously unknown images as pornographic, automated methods need a way of scanning the image for hints of adult content: A common approach focuses on detecting skin in pictures by analysing colour histograms or investigating shapes (Xu et al., 2020). While pornography certainly features an above-average display of skin, skin alone does not function as a reliable classifier for pornography. These approaches are hence error-prone to produce false-positives like beach photos and false-negatives, such as leather/latex porn. Additionally, these models struggle with different skin tones and objects shaded in "skin tone" (Nian et al., 2016). A second approach aims at extracting local features such as shape and texture, and subsequently applies probabilistic models onto those to infer higher-order features (Xu et al., 2020; Karamizadeh and Arabsorkhi, 2018). Yet, these approaches are outperformed by applications of *convolutional neural networks*, a subtype of artifical neural networks, effective for the analysis of images (Xu et al., 2020; Nian et al., 2016; Karamizadeh and Arabsorkhi, 2018). Convolutional neural networks are widely applied in commercial software for the visual analysis of images, such as GoogLeNet or Facebook's DeepFace (Alake, 23.12.2020; Voulodimos et al., 2018).

In the following, I shortly digress to explain the procedure underlying convolutional neural networks. Images on computers are represented as digits for each pixel. In coloured images, each pixel is represented by three values, one for each colour channel (red, blue,

green). These values indicate the intensity of the respective colour at this location. Convolutional neural networks help turn this low level information into more abstract and meaningful attributes, e.g. 'Is this photo showing a cat or a dog?', 'How many people are on this picture?', or eventually, 'Does this image portray illicit content?'. Pornography detection can be formulated as a (binary) classification task (e.g. pornographic content/non-pornographic content), or as a (logistic) regression task, resulting in an estimated probability that the given image contains pornographic content. To obtain this higher-level information, images are analysed by small blocks of neighbouring pixels. Each block is scanned for the occurrence of a specific feature, such as a vertical edge. This operation called *convolution* produces a new image that can be conceived as a map, indicating where in the original picture the feature occurred. Convolutional neural networks stack multiple of these maps on top of each other, hence built upon the features that previous layers have already extracted. This way, complex abstract information can be extracted just from different pixel shades (Goodfellow et al., 2016, p. 6). The extraction of useful features hinges on the right adjustment of the parameters. For deep neural networks with often millions of parameters, this is achieved in a process of iterative fine-tuning using large amounts of data. Adjusting a model's parameters to improve its performance is referred to as *training*. One method for training models is *supervised learning* which requires previously labeled data. In the process of training, the parameters are iteratively updated to reduce the error between the model's predictions and the correct labels. In the end of this training, this error should be minimized, so that the model is able to generate appropriate predictions for unseen images with a high probability.

This simplified procedure illustrates that automated models like convolutional neural networks do not have a *semantic* understanding of pornography, nor any other concept. The distinctions they perform stem solely from the information handed over to the model from the data, or more precisely, the humans that labeled the images in the training data set. The performance of pornography detection models thus depends on the cultural assumptions prevalent in the training data. Yet, also models that are not trained on previously labeled data reflect the notions of the people involved in their making. In *unsupervised* approaches, models aim to infer a distinction or classification solely based on the inherent characteristics of the given images. However, when evaluating their performance, these models are still measured against the expectations of their developers and commissioners. A model that outputs a divergent classification of pornography would be further fine-tuned until its predictions are in line with the developers' notion of pornography.

In a meta-analysis of publications that propose automated approaches toward pornography detection, Gehl et al. (2017) observe that the conceptions present in these models largely reflect a normative understanding of bodies and sexuality. They argue that the 'overwhelmingly straight male population' of computer scientists inscribes their narrow and conservative conception of sexuality into these models. With the application of these

models, the narrow set of assumptions they are based on is further disseminated, reinforcing gender and sexual inequalities.

3.1.4 The "Tumblr Porn Ban"

To illustrate the mechanisms of content moderation systems I use the example of the so-called "Tumblr Porn Ban", in 2018. This event constituted a large-scale and drastic shift in moderation policy of a popular platform, and elicited a public outcry from its shocked and disappointed user base (e.g. Pettis, 2020). Due to the vocal public criticism the Tumblr Porn Ban gained an infamous reputation in pop culture. In the past few years since its implementation it has been the subject of research in communication studies, porn studies, and queer theory, thoroughly documenting its wide-scale effects (e.g Bronstein, 2020; Pettis, 2020; Ashley, 2019; Engelberg and Needham, 2019).

Tumblr is a social media platform that allows users to create blogs, so-called *tumblrs* filled with multimedia content—images, videos, text, ect. The 'dashboard' start page arranges recent posts from the users one follows into personalized feeds, curated based on the user's interests (e.g. Cho, 2015; Pettis, 2020). The site was known for its laissez-faire stance towards pornography moderation, not restricting the distribution of sexually explicit contents. This made it a haven for explorations outside of the mainstream. As Ashley (2019) recalls her active time on Tumblr:

> Tumblr was the first place I'd experienced porn presented as having aesthetic and cultural value. It was out in the open. It allowed you to become a collector of your own desires, displaying them and celebrating them proudly, rather than having them spoon fed by a tube-site algorithm.

The online space of Tumblr was particularly popular for young queer folk. One reason for Tumblr's popularity in the queer community might be the fact that it relies on relative anonymity (Cho, 2015; Pettis, 2020). Other than on Facebook where users are suggested to identify with their real name and connect their profile to their real-life relatives and friends, on Tumblr, one is protected by a pseudonym. Users who have to keep their queer desires and identities veiled in their offline lives found a space to explore without unwanted consequences on Tumblr (Cho, 2015).

In December 2018, Tumblr announced a change in their community guidelines, effecting the removal of any type of adult content. From December 17th 2018 on, 'photos, videos, or GIFs that show real-life human genitals or female-presenting nipples, and any content [...] that depicts sex acts' (Tumblr, 16.07.2020) would no longer be visible to other users, only to the person posting it (Tumblr, 09.01.2021). Tumblr pledged to make exceptions for nudity in non-sexual contexts, such as 'exposed female-presenting nipples in connection with breastfeeding, birth or after-birth moments, and health-related situations, such as post-mastectomy or gender confirmation surgery.' Further, '[w]ritten content such as

erotica, nudity related to political or newsworthy speech, and nudity found in art, such as sculptures and illustrations' (Tumblr Support, 2018) should also continue to be permitted on the site.

Aware of the challenge of making nuanced decisions, Tumblr stated it will be using 'a mix of machine-learning classification and human moderation' (Tumblr Support, 2018). As the exact composition of the content moderation system used by Tumblr is publicly unknown, it cannot be stated with certainty to which extent flagging of explicit content was performed by ML models, and to which extent by human moderators. However, the sheer pace at which images were taken off the platform, resulting in a 'purge' (Engelberg and Needham, 2019) of content over the subsequent weeks, strongly suggests the substantial role of automated ML approaches in implementing the ban. Moreover, especially in the beginning, the flagging procedure exhibited a high error rate, frequently flagging content that quite obviously appears non-sexual to the human eye (Matsakis, 06.12.2018). This indicates that Tumblr heavily relied on automated models to enforce the ban, owing the erroneous outcomes to insufficient training, at least at the beginning of its application (Pilipets and Paasonen, 2020). The short period between the announcement of the ban and its implementation further fortifies this presupposition.

Tumblr's reasons for implementing the ban were multifold, ranging from changes in legislation that make platforms legally viable for sex trafficking, allegations of hosting instances of child pornography to a change in management and advertising strategy (Bronstein, 2020; Pettis, 2020). I return to these reasons in the sections 4.4.1 and 5 where I embed them into their larger context in the constellation of power.

The Tumblr Porn Ban constitutes a decisive example of how sexually progressive values on an online platform were restrained, and how ML-aided content moderation was implicated in propagating a hegemonic sexual norm. In section 4.4.1, I explore in more depth how this ban affected various marginalized social groups to exemplify the ways through which power operates to reproduce normative alignments.

3.2 Pornographic deepfakes

Turning to the second application of ML for pornography, this section explains the phenomenon of deepfakes, their technical implementation and their most common usage, the creation of unconsensual pornography.

Deepfakes, also sometimes referred to as 'faceswap', is a technology that allows to insert other faces onto bodies in video material. As the name suggests, deepfakes are founded on DL technology which allow users to create realistic videos, making it look like a person is doing or saying something they might never have said or done. The videos are usually prepared by giving the algorithm photos of the person and having it automatically

detect corresponding facial features to swap. Manipulation in real time is also possible (Pinscreen, 2020).

As a brief introduction to the technology behind, deepfakes are facilitated by 'generative adversarial networks', in short GAN. GANs employ two separate models that are trained at the same time against each other. The first part, the so-called 'generator' receives a large amount of image data (for instance of human faces) and is tasked to create novel images similar to those. The second model, called the 'discriminator', is optimized to distinguish fake images from real ones (Goodfellow et al., 2016, p. 690). To achieve this, both models rely on convolutional layers, as described in section 3.1.3. As the name 'adversarial' suggests, generator and discriminator are trained simultaneously against each other, thus allowing both to continuously increase their performance: The better the discriminator gets at detecting the fake images the more the generator is forced to create more realistic images.

In the past two years, deepfakes gained a considerable amount of public attention in particular because of their risk of political misinformation and the epistemic threat they pose to democracies (e.g. Fallis, 2020). However, an investigation of the deepfake landscape online found that political deepfakes only represent a tiny fraction of use cases. The most common use of deepfakes is pornography. In 2019, the company Sensity.ai detected more than 14.000 deepfake videos online, 96% of which were pornographic (Ajder et al., 2019). And the scene is thriving. Six months later, in June 2020, the number of total deepfakes had already more than tripled, reaching almost 50.000 videos (Ajder, 2020). The report further observes that these pornographic deepfakes exclusively target femme-presenting people. Male-presenting figures dominate the small fraction of non-pornographic deepfakes (Ajder et al., 2019) that often portray politicians or actors, such as Nicolas Cage and Donald Trump who are popular characters of non-pornographic deepfakes (Newton and Stanfill, 2020; Winter and Salter, 2020). Deepfake porn videos frequently depict famous actresses, singers or social media personalities (Ajder et al., 2019). However, also regular FLINTA* become targeted, often initiated by malicious ex-partners[7] (Winter and Salter, 2020; Alptraum, 15.01.2020). Researchers thus consider deepfake porn a gendered phenomenon that is exercised by (presumably heterosexual and cisgendered) men to target FLINTA* (Newton and Stanfill, 2020; Winter and Salter, 2020). The phenomenon kicked off in 2017 on Reddit, where a user named 'deepfakes' posted porn videos in which porn performers' faces were replaced by celebrity's faces. The subreddit was banned shortly after by the platform's administration due to its dissemination of involuntary pornography (Jacoby, 09.12.2019).

[7]Due to cases like these, deepfake porn is also called 'revenge porn'. However, this term is problematic, as Dunn (2020) describes: 'The term "revenge" suggests that the person in the images was deserving of the abusive disclosure of their images, and the term "pornography" suggests that the images may be legitimately used by unintended audiences for sexual purposes.' A more appropriate term is image-based sexualized violence.

One of the earlier applications of pornographic deepfakes targeted the Indian investigative journalist Rana Ayyub. Ayyub had uncovered the complicity of Indian government officials in riots and murders, and has since then been the target of sexist and anti-muslim hatespeech on the internet (Ayyub, 22.05.2018). In April 2018, a pornographic deepfake video featuring her was distributed online. Seeing the video and its rapid distribution caused her enormous pain (Citron, 2019) and intensified the hatred and misogyny she was subjected to (Ayyub, 21.11.2018). She reflects that the deepfake video aimed at crushing her credibility as a public figure and silencing her political criticism (Ayyub, 22.05.2018). And it was, at least temporarily, successful:

> From the day the video was published, I have not been the same person. I used to be very opinionated, now I'm much more cautious about what I post online. I've self-censored quite a bit out of necessity. Now I don't post anything on Facebook. I'm constantly thinking what if someone does something to me again. I'm someone who is very outspoken so to go from that to this person has been a big change. I always thought no one could harm me or intimidate me, but this incident really affected me in a way that I would never have anticipated. (Ayyub, 21.11.2018)

Ayyub's testimony illustrates the sinister effects of pornographic deepfakes. Individuals targeted in these videos are harmed in their autonomy and experience 'severe emotional distress' (Citron, 2019) that can have long-term effects such as 'anxiety, physical illness, and job loss' (Maddocks, 2020). The online distribution of the material is often accompanied by a flood of hateful messages on social networks, including slut-shaming, and rape and death threats, further hurting the victims' reputations and intensifying their suffering (Citron, 2019; Ayyub, 22.05.2018). These threats are not confined to online spaces since perpetrators commonly publish victims' real names, addresses and phone numbers alongside the footage—a method also known as 'doxing'[8] (Dunn, 2020).

3.3 Justification for this selection

Besides content moderation systems and deepfakes, also other current examples of ML systems entice an investigation into their ability to reinforce hegemonic sexual norms. One particularly alarming instance is recent research into predicting a person's sexual orientation from a photo of their face using deep neural networks (Wang and Kosinski, 2018). In this work, however, I limit my considerations to the phenomenon of pornography, as one highly relevant, but often scholarly neglected cultural practice. In the examples of content moderation systems and deepfakes, one can observe how ML penetrates the field of pornography, influencing both the creation as well as the deletion of pornographic artefacts.

[8] alternatively 'doxxing', stemming from documents

Analysing these two applications in conjunction not only sheds light onto the field from these two different angles, but further allows for interesting synergies. Both technologies cannot only be seen as isolated phenomena, but are entangled with each other: Central to the public debate around deepfakes is the call to remove them from porn sites like PornHub (e.g. Wagner and Blewer, 2019), making them a matter of content moderation systems. Considering the mechanisms of content moderation systems in conjunction with deepfakes allows a more profound understanding of their interwoven workings.

4 Power

In the previous section I have introduced two example technologies situated in the intersection of pornography and ML. Before I investigate the concrete norms reinforced by content moderation systems and pornographic deepfakes and in which way they are aligned with a hegemonic form of sexuality, I elaborate more generally on the mechanisms through which power operates, incorporating insights from the works of Michel Foucault and Sara Ahmed.

When thinking about power in the two examples, an observation comes to mind: In content moderation systems, sexually explicit content is *repressed*, in the way that it is not allowed to be seen or engaged with on the platform. At the same time, however, content moderation systems also *produce* a certain order, a distinction between what counts as "normal" and what is considered "illicit". In pornographic deepfakes, the depicted people suffer severely under the distribution of the videos (Citron, 2019). In examples like the case of Rana Ayyub, who was forced to cease her critical reporting, the *repressive* manner of power comes to light. Yet, through the availability of such videos, also new conceptions, practices and desires are formed, which illustrates that power operates in a *productive* manner simultaneously.

This observation draws links to the concept of power in the work of Michel Foucault who has created an extensive theoretical body around the concept of power. To fit the scope of this work, I limit my consideration to a few aspects that I consider most relevant for this application. Most centrally, an understanding of power as *relational*: Foucault contests the belief that power can be possessed as a good; for him power can neither be held by an individual or a group, nor obtained or taken away. Power, instead, is 'exercised from innumerable points, in the interplay of nonegalitarian and mobile relations' (Foucault, 1978, p. 94). Accordingly, it cannot be envisioned as a unidirectional top-down process, but rather as arising from 'below'—from the microlevel of everyday interactions (Foucault, 1978, p. 94). Conceiving power this way shifts the focus away from questions regarding the *ownership* of power toward the *functioning* of power (Mühlhoff, 2018, p. 260). This understanding is useful when observing the dynamics within and across platforms where no single agents exerting power over others can clearly be identified as the "sources" of

power. Further, making power intelligible as a bottom-up process is a helpful foundation for developing an understanding of how power can arise from the masses and accumulate into hegemonic norms on platforms. Moreover, Foucault emphasizes that power does not operate merely negatively, by repressing what is undesired, but also contains a *productive* side. I elaborate more thoroughly on this productive mode of power in the subsequent section, where I illustrate its involvement in establishing norms.

Aligned with the focus of this work, it is of great interest to understand how the mechanism of power to form and enforce norms surrounding gender and sexuality. This question is also of central matter to queer theory, a field of study that critically examines normativity and deviance, especially in the realm of sexuality and gender (Spargo, 1999). The insights of queer theory are thus particularly apt for this investigation. To incorporate a queer perspective, I subsequently introduce Sara Ahmed's concept of *straightening devices* that explores the (re-)production of norms like heteronormativity from a bodily, experiental level. Yet, as a foundation, I begin with elaborating Foucault's notion of productivity of power.

4.1 Productivity of power in Foucault

In his genealogy of the prison, *Discipline and Punish*, Foucault (2012) traces the emergence of productive power in parallel to the erection of modern disciplinary societies. He observes how in medieval societies, rulers overtly exerted punishment over non-compliant citizens, often in public spectacles as a means of maintaining their status. Yet, over the course of the 19$^{\text{th}}$ century the focus shifted from punitive toward disciplinary strategies. These are characterized by the establishment of norms that lay out a notion of what it means to be an ideal citizen. Institutions like the military, hospital and school played crucial roles in propagating these norms and developing the discipline necessary to achieve them (Fink-Eitel, 1989). Governing could then operate more opaquely, avoiding public performances of punishment (Foucault, 2012, p. 9), and more effectively, since the drive to accomplish norms would come from subjects themselves, without the need of external repression. Power, hence, could no longer be considered only negatively, in terms of inhibiting or suppressing, but as positively, producing new behaviours, pleasures and knowledge:

> What gives power its hold, what makes it accepted, is quite simply the fact that [...] it induces pleasure, it forms knowledge, it produces discourse; it must be considered as a productive network which runs through the entire social body much more than as a negative instance whose function is repression. (Foucault, 1979, p. 32)

Although in statements like these Foucault highlights the productive side of power, the ability of power to act repressively is not excluded. Instead, the characteristic of power can be understood as a *regulating* force which uses technologies and strategies to nurture certain qualities and fend off others (Jäckle, 2014). Accordingly, the elements of power

present in deepfakes and content moderation systems contain both modes simultaneously, as I explore in more depth in the following sections. Through the production of new knowledge and behaviours, power, eventually, is also implicated in forming subjects, as Johanna Oksala describes:

> By claiming that power relations are productive of forms of the subject, Foucault does thus not simply suggest that individuals are produced as subjects just as cars are produced from various materials in a factory. Rather, we must understand the subject to be intrinsically entangled with power and knowledge. Power/knowledge network constitutes the subject in the sense of forming the grid of intelligibility for its actions, intentions, desires and motivations. (Oksala, 2005, p. 95)

One can take away that the productive element of power operates through the construction of an all-encompassing system of beliefs and norms, a *grid* that constitutes the basis for explaining what we think, want and do. I want to pick up on Oksala's 'grid of intelligibility' and relate it to a concept by queer-feminist philosopher Sara Ahmed that studies norm production from a phenomenological perspective, the concept of 'straightening devices'.

4.2 Ahmed's straightening devices

In her book *Queer Phenomenology* Ahmed explores a phenomenological account of *orientations*. With the focus on orientations Ahmed invites us to rethink the emergence of normativity from the experiential perspective of bodies and space: how and where bodies are situated in space, how they relate, 'tend' (Ahmed, 2006, p. 51) toward other objects and bodies and what is (un)available for a certain body given its orientation in space (Ahmed, 2006, p. 3, pp. 50–55). All these conditions shape and re-arrange the space and can lead to higher structures emerging on it. Ahmed refers to the latter as 'lines' (Ahmed, 2006, p. 12) which can be understood as the dominant tracks that guide bodies.

'Straightening devices' tamper with these orientations and try to bring bodies into alignment with the normative line (Ahmed, 2006, p. 72). One example for straightening devices that Ahmed gives is 'compulsory heterosexuality' (Ahmed, 2006, p. 23, pp. 84–91): compulsory heterosexuality imposes an orientation onto bodies that naturalizes the desire toward the "other" sex. It constructs males and females as opposite sexes who complement, complete each other and relies on the narrative of penises and vaginas as 'made for each other' (Ahmed, 2006, p. 71, p. 85). The imposed orientation (that is nonetheless perceived as natural), from female bodies to male bodies and vice versa, modifies what bodies tend toward but also what bodies can reach. For instance, it determines who is reachable as a potential object of love and desire (Ahmed, 2006, p. 95). This orientation is further enforced in institutions like marriage (Ahmed, 2006, p. 84). Hence, heterosexuality arranges objects in a certain way and lays out a delineated space for action, a

line. This line can be considered the normative: what is not straightly oriented the same way is construed as deviant, abnormal and *queer* (Ahmed, 2006, p. 72). Consequently, straightening devices play a central role in bringing about normativity, as Ahmed writes:

> [T]he normative can be considered an effect of the repetition of bodily actions over time, which produces what we can call the bodily horizon, a space for action, *which puts some objects and not others in reach.* [...] [T]his alignment depends on straightening devices that keep things in line, in part by 'holding' things in place. (Ahmed, 2006, p. 66, original emphasis)

Straightening devices thus not only guide bodies on the level of sexual attraction, but elucidate more profoundly the operation of power, how norms arise and *what gives them their hold*. Ahmed underscores two aspects of straightening devices that are constitutive of normativity: Firstly, they determine what is available for certain bodies and this way create a bodily horizon. And secondly, they enforce this bodily horizon through repetition. Content moderation is a helpful example to illustrate these mechanisms of straightening devices to produce and enforce normative alignments.

Content moderation in its very nature is constructed to 'put some objects and not others in reach'. It aims to keep objects out of reach that do not comply with platform guidelines, such as posts that are portraying violence, nudity, pornography, etc. By doing so, it performs a definition of what can be expressed or seen on a given platform. Content moderation systems, thus, 'keep in line' the sayable and visible.

I would like to accentuate specifically the *virtual/digital* horizon as a part of the 'bodily horizon'. Although Ahmed herself does not mention the digital realm in *Queer Phenomenology*, the 'bodily horizon' can be understood in a way that does not only refer to the concrete, physical body. Instead, the 'space for action' transcends the literal meaning and can also refer to abstract phenomena, in the sense that a certain legislation for instance can also extend or shrink a bodily horizon. Nonetheless, by accentuating the term 'digital horizon' I want to highlight one particularly powerful part of our bodily horizon and its extensive ability to create normative alignments.

Digital media, in particular social media platforms, play a substantial role in mediating our perception and shaping our experience of the world. With the help of digital media our experiential horizon has vastly expanded, allowing us to visually and auditorily engage with events taking place thousands of kilometers away, as well as events that took place weeks, years, decades ago. Time spent online is increasing and shifts bodily interaction and perception toward digitally mediated forms of correspondence.[9] However, everything we can perceive through social media platforms has been automatically filtered, ranked

[9]This shift to digitally mediated forms of interaction becomes particularly palpable now, in times of a pandemic, where digital media make up a substantial part of our access to the world.

and preselected for our eyes and ears. These selective representations are algorithmically compiled into the linear structure of endless scroll feeds, keeping *in line* what is 'on-line'. The mediation offered on online platforms, thus, profoundly shapes our horizon, altering what is available and what is not.

This powerful digital horizon is then consolidated through repetition. Again, the purpose of automated content moderation systems is precisely to scale up the number of decisions that can be taken (see also Cobbe, 2020). Innumerable times each second and all over the globe, moderation decisions are made, re-inscribing a norm into the space. They determine not only what is visible for us, but also delineate what can come to mind at all to us, and what is normal, right or illicit for us.

Ahmed's conception of normativity and straightening devices emphasizes that norms and power operate on an immediate, bodily level. This way, she underscores that power is an all-encompassing grid which incorporates, and guides bodies. Further, Ahmed elucidates that the normative is a matter of *alignments*. In other words, the normative arises as a result of a specific *arrangement* of various other elements. For instance, the presence of narratives and discourses that confirm and justify the normative direction, such as the naturalization of penises and vaginas as "made for each other". Straightening devices like content moderation also contribute through continuous repetition to the propagation of the normative line and thus become components of the constellation.

With this observation in mind, I return to Foucault and introduce a final aspect of power, the concept of the apparatus.

4.3 Foucault's apparatus

To inspect how power operates through complex arrangements of various elements, Foucault offers the concept of the apparatus (*dispositif*). Such a 'heterogeneous ensemble' achieves a specific way of operating through the interaction[10] of its components (Foucault, 2008, p. 194). These components can be both material or discursive, such as 'discourses, institutions, [...] laws, administrative measures, [...] moral and philanthropic propositions' (Foucault, 2008, p. 194). More than the elements themselves, it is the relations between them that matter, 'the network that connects and disconnects these elements, and determines the distribution of power and knowledge' (Nikolić, 2017, p. 133). This network and the specific arrangement of its parts embodies a certain 'inclination, tendency' (Nikolić, 2017, p. 133), as the French term 'dispositif' suggests. An apparatus thus performs a 'strategic function' (Foucault, 2008, p. 195), which can be understood as an operation 'which fixes, reproduces, multiplies and accentuates existing relations of forces' (Foucault, 2008, p. 203).

[10] Or *intra*-action, a term that highlights that elements are not separate entities, but in their acting-together form a new entity (Barad, 2007).

In the following, I demonstrate how the ML systems of content moderation and pornographic deepfakes constitute powerful apparatuses, incorporating ML models, pornography, platforms, users, legislation and economic imperatives. Additionally, in exemplifying their capacity to promote a hegemonic sexuality, I expose their ability to fix and reproduce 'existing relations of forces' (Foucault, 2008, p. 203)', which can be identified as their *strategy*.

4.4 Enforcing hegemonic sexual norms

After introducing relevant concepts of power and norm production, in this section, I turn to the concrete effects that arise from content moderation systems and pornographic deepfakes. I consider how the ML systems as apparatuses are implicated in inscribing hegemonic norms into discourses, gendered identities, sexual practices and material conditions. Put more specifically, for both applications I give examples that illustrate how heteronormativity, hegemonic conceptions of femininity and masculinity as well as hegemonic material conditions are echoed or amplified.

4.4.1 Content moderation systems

I begin with content moderation systems and return to the example of the Tumblr Porn Ban. As mentioned in section 3.1.4, the Tumblr Porn Ban has been thoroughly debated by the public and investigated by scholars. Therefore, its wide array of effects is well-documented, making it a beneficial aid for contextualizing the impact of content moderation systems in enforcing a hegemonic sexuality.

With the implementation of the Tumblr Porn Ban in 2018, an important space for heterogeneous pleasures, desires and bodies was eradicated. Tumblr was a popular place for queer users (Cho, 2015) and exhibited adult content outside of the heteronormative and monosexual mainstream. On Tumblr, queer pornography could appear interwoven with other content, such as design or landscape images. This heterogeneous 'rhizomatic' feed was a counterdraft to other social media platforms which ban explicit sexual content from their sites (Ashley, 2019).[11] Tumblr's design also differed from the taxonomies presented on most porn tubesites where queer porn is kept strictly separated from straight porn (Engelberg and Needham, 2019). Instead of hiding queer porn from view or confining it to the platform's shady margins, it prominently entered the main stage usually reserved for straight sexuality. The multitude of pleasures and desires that was accepted on Tumblr challenged existing rigid sexual identity categories and allowed subjects to rethink their own desires more holistically. Moreover, the editorial practices that composed posts into carefully curated feeds were able to perform 'hermeneutic shift[s]' (Engelberg and Needham, 2019) that subverted dominant straight readings. For instance, by re-arranging video

[11] For an account of the rhizomatic character of Tumblr see Cho (2015).

sequences of accidental male ejaculation on other straight men during group sex, 'viewers could find unintended queer pleasures' (Engelberg and Needham, 2019). Reframing straight porn in this way to give it a new queer meaning called into question the authority of hegemonic sexual practices and readings. With the implementation of the ban, such subversive reframings and explorative potentials were curbed. In effects like these, the ML-aided Tumblr Porn Ban can be read as an enaction of the strategy to restore and propagate identity categories and practises that support the hegemonic sexual order.

A similar mechanism can be observed for restoring hegemonic body norms: Tumblr had offered representation to bodies which are marginalized on mainstream platforms. For instance, various blogs explicitly centred 'small' or 'average-sized' and unerect penises as objects of desire. Engelberg and Needham (2019) read this practice as an act of resistance against hegemonic body norms that equate penis size with masculinity. Penises, or more precisely the erect phallus, have been identified as a crucial element of hegemonic masculinities (e.g. Bollas, 2021) and are also fundamental components of hegemonic sexuality. Firstly, in the hegemonic reading, the erect penis functions as a symbol of dominance, power and virility, a framing that becomes even more evident in its negated form: A penis that is unable to get hard is signified through the term *impotence*, derived from Latin for 'without power' (Potts, 2000). Secondly, the act of penile penetration plays a crucial role in hegemonic sexuality, where penile-vaginal penetrative sex is constructed as the ultimate and only form of sex (see also Bollas, 2021). Other sexual actions are commonly considered merely "foreplay", the name discursively enforcing that what is *before* sex cannot be sex itself. This results in a hierarchization of sexual practices that posits heterosexuality at its top and subjugates other sexual practices, like lesbian or gay sex. The subversive reframing of penises as soft, delicate and vulnerable objects (hooks, 01.07.1999) in fact 'feminize[s]' men (Potts, 2000). Such a reading would reject the notation of power ascribed to the phallus and would thus erode the difference hegemonic masculinity constructs between sexes to legitimize the privilege of cis men. Consequently, in this example, the implementation of the porn ban through content moderation systems can be interpreted as power operating in a regulating manner to keep intact readings of bodies that are necessary to legitimize—and hence proliferate—existing power relations between sexes.

Further, the communities on Tumblr were previously welcoming to trans or genderqueer people, giving them a space to talk about common struggles or topics of pleasure (Bronstein, 2020). Making bodies outside of the cis-sexist norm visible serves as an invaluable epistemic resource for people of indeterminate gender as it produces visual artefacts of living and loving with a norm deviant body (Bronstein, 2020). This representation challenged hegemonic sexuality which either refuses the right of these bodies to be represented or denies their existence altogether. After the ban, many of these artefacts were no longer available. Although Tumblr stated to allow medical-related trans content like gender confirming surgeries, this often proved to be wrong with trans users reporting that their

content documenting genital surgeries was nonetheless flagged as 'adult' (Bronstein, 2020). Yet, even without false classification, the distinction between medical content as legitimate and other content relating trans sexuality as 'illegitimate and pornographic' constitutes a problematic reduction of trans people to medical discourses and the state of their genitals (Bronstein, 2020). With the deletion of porn from Tumblr, depictions of trans sexuality are confined to mainstream porn platforms, where trans bodies are often portrayed in a fetishized and exoticizing manner (Bronstein, 2020). Their pleasure is playing an inferior role, making it rather porn *about* than *for* trans people. Through the reduced visibility of trans/genderqueer bodies, the Tumblr Porn Ban ultimately re-affirmed the cis-binary body norm. It further restricted trans self-determined sexual expression, trapping them in the role of consumable objects that does not challenge the hegemonic norm.

The Tumblr Porn Ban specifically hurt sex workers who used the site to advertise their business. Recent US legislations[12] that aim at preventing sexual exploitation and trafficking make platforms legally liable for sex work that is promoted on their sites (Bronstein, 2020). These legislations have also been identified as one of the main drivers of the Tumblr Porn Ban (Cyboid, 2018). Yet, what sounds like a noble cause has devastating consequences for sex workers: As a result, platforms have erased ways for sex workers to promote their business online. Without the option to advertise online, sex workers lose their financial basis and are pushed into more precarious labor conditions. Sex workers with direct client contact are pushed away from the safety of online communication into the streets or the dependencies of intermediaries ("pimps") (Tripp, 2019; Singh, 2019). Video performers lose their viewerbase and their traffic is re-directed toward mainstream porn sites like PornHub which predominantly host pirated videos that do not adequately compensate performers (Bronstein, 2020). While legislation and ressentiments against sex workers are certainly not novel phenomena, content moderation automated through AI takes control over sex workers to a new level. Automated models can scan through whole platforms in a pace that human moderation could not keep up with. Power, in this example, is present in its repressive character, as certain actions, namely the selling of sexual services, are inhibited. At the same time, however, anti-sex-work content moderation is also productively involved in forming a certain reality and arranging material circumstances in a way that favors certain groups, production methods and values. Undermining conditions for self-organized and self-determined sex-work benefits large established porn companies which rather produce profitably for the mainstream than to create innovative, inclusive and liberatory content for marginalized communities (Ashley, 2019). Automated content moderation thus becomes a factor that proliferates a homogeneous sexual online landscape and control over sex workers.

I have focused on Tumblr exemplarily, yet similar mechanisms of promoting a hegemonic sexuality can also be found on other platforms. On Instagram, content moder-

[12]Stop Enabling Sex Traffickers Act (SESTA) and Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA)

ation systems deleted pictures containing femme-presenting nipples, menstrual blood or 'unshaven bikini lines' (Faust, 2017). Eradicating instances of these 'features of basic physiology' (Faust, 2017) as pornographic or otherwise illicit content illustrates the sexualization or "pornofication" of femme bodies that cause a topless femme body to be considered pornographic, whereas a bare male chest can easily prevail on the platform. Moderation policies like these promote the hegemonic understanding that femme bodies mainly serve for objectification and hinder alternative, self-determined depictions of femme bodies.

On YouTube, certain content which is not deemed appropriate for all audiences is restricted in its visibility (blog.youtube, 20.03.2017).[13] A disproportionate amount of these restricted videos comprised content related to LGBTQ* topics (Southerton et al., 2020; Neonfiona, 16.03.2017). Although in effect, the visibility is reduced only slightly, the material consequences for creators are substantial: Since the earnings of producers are dependent on the number of views and advertisers are more reluctant to place ads alongside restricted content, producers of LGBTQ*-related content are materially disadvantaged in comparison to creators who cover similar topics from a straight perspective (Southerton et al., 2020). Policies like these that take the hegemonic norm for orientation undermine the production of content that displays perspectives outside of the mainstream. Further, Southerton et al. (2020) have observed how through the selective restriction of LGBTQ* content, YouTube performs a distinction between "good" and "bad" queer subjects. The *good* queer subject is symbolized in coming outs or same-sex weddings and constructed as 'the one who participates in and seeks validation from heteronormative social institutions like marriage and the family' (Southerton et al., 2020, p. 10). By contrast, the *bad* queer subject is characterized as being overly sexual, its open display of norm-deviant sexual behaviours is perceived as a threat to other users (Southerton et al., 2020). This distinction and the scarcity of online coverage it entails also drive a further divide into the queer community, leading some to rather dissociate themselves from the "socially unacceptable" forms of queer expression than to show solidarity with those being policed by moderation guidelines (Ashley, 2019). Hence, it becomes evident that content moderation not only exhibits a bias against LGBTQ* content, but is also implicated in the construction of a queer sexual subjectivity that is docile to and allied with the hegemonic norm.

In these examples, pornography-centred automated content moderation is used to impede the visibility of queer perspectives, deviant bodies, pleasures and subversive reframings. The restraining of these forms of expression reveals the *repressing* manner of power. At the same time, power also operates *productively* in the given examples. Through the selective erasure of deviant pleasures, practices, perspectives and bodies, those that are allowed to remain are asserted as acceptable and advance the established normative sexual conceptions. This way, content moderation becomes involved in determining what is

[13]The 'restricted mode' is an opt-in option, conceived for schools, libraries, ect. With the restricted mode deactivated, the videos are visible just like other videos.

available, and what, as well as who, is considered normal. The repetition and scale of its decisions, increased through automation, contribute to legitimizing and upholding the dominance and privileges of straight cis bodies. Further, through making some sexual desires invisible and promoting others, content moderation can affect how subjects make sense of their desires, and align their perceptions with the conceptions fed on mainstream social media and porn platforms. By making users internalize dominant notions of sexuality and identify themselves with it, these norms can be propagated more effectively, without requiring external repression. Hegemonic sexuality is solidified on a material level, too. Through financially privileging straight perspectives over queer ones, content moderation systems have an influence on what type of content is lucrative or even affordable to be produced and consumed in the future. Consequently, content moderation systems can fortify conditions favorable to the advancement of the hegemonic sexuality and proliferate the norm more long-lastingly.

It cannot be concluded that therefore content moderation systems inevitably serve a hegemonic norm. The examples I have given by no means represent an exhaustive analysis of the field of pornography-centred content moderation. There are certainly counterexamples, for instance content moderation that is applied for removing non-consensual pornography from platforms. Without giving a comprehensive review of the field, these examples nonetheless offer a glimpse on the forces that are at work in contemporary content moderation. They illustrate how content moderation is applied, for which purposes, and with what effects, whose voices are heard in the process and who is disregarded as the "collateral damage" of moderation. This approach of shifting the view to concrete *effects* of content moderation is in line with Foucault's understanding of how power is distributed over heterogeneous composites like ML systems (Jäckle, 2014).

4.4.2 Pornographic deepfakes

In the following, I turn to pornographic deepfakes and illustrate how in this application, hegemonic sexual norms are inscribed through conceptions of masculinity and femininity, (sexual) practices and material conditions.

As mentioned in section 3.2, deepfake porn specifically targets femme-presenting people and is commonly created without the consent of the person depicted. The assumption that the agency, bodily autonomy of FLINTA* can be subordinated is thus a foundational premise of deepfake porn. Yet, deepfake porn not only incorporates and reflects existing sexist conceptions, but also further fuels them. The availability of deepfake porn contributes to an understanding in which it is *normal* to disrespect the autonomy, boundaries and consent of femme bodies. As a result, femme bodies are rendered as passive objects, free at disposal for the enjoyment of others.

Recalling the example of the investigative journalist Rana Ayyub, one can note how pornographic deepfakes achieve to suppress certain qualities that are undesired to the

hegemonic norm, such as being a politically uncomfortable woman. The terror of deepfake porn can selectively punish FLINTA* that step out of the role intended for them by hegemony. Strikingly, deepfake porn particularly targets FLINTA* which wield their agency, e.g. when voicing political criticism or ending an unfavorable relationship. Yet, the effect reaches far beyond the people who are directly targeted. Deepfake porn contributes to a climate of violence which can give rise to new behaviours and conceptions, elicited by finding mechanisms of coping with it. In a study of Indian women, Gurumurthy et al. reported how a climate of gender-based violence online led participants to alter their profiles in compliance with traditional gender norms, 'to cultivate a "good girl" image and survive online using the "language and expectations of patriarchal logics"' (Gurumurthy et al., p. 21). Further, subjects exposed to digital violence like deepfakes can exhibit tendencies to normalize it, accustom to it or blame themselves (Dunn, 2020). These observations illustrate that the threat of gendered violence exerted by pornographic deepfakes not only operates repressively, by inhibiting norm-deviating behaviours. It simultaneously produces certain behaviours and subjects which are docile to the hegemonic gendered norm. Specifically mechanisms of having individuals internalize and naturalize the violence they are subjected to allows power to operate more effectively and opaquely (Foucault, 2012, p. 9). Put in the words of Sara Ahmed, deepfake porn functions as a straightening device which 'holds in place' a notion of subordinated femininity and pushes deviating bodies back 'in line'.

The misogyny prevalent in deepfake porn also serves to perpetuate the existing hegemonic hierarchization of genders that posits masculinity above femininity. But not all men benefit equally from this hierarchization. Toxic geek masculinity is a form of masculinity that is subordinated to the hegemonic archetype and that has been associated with the online spaces where deepfake porn is developed and distributed (Newton and Stanfill, 2020). Men attributed to this type of masculinity do not meet the standards of hegemonic masculinity, for instance because they are not considered attractive or successful[14] enough. Toxic geek masculinity is a common media trope and is, among others, embodied in the main characters of the series 'The Big Bang Theory' (McIntosh, 2017a,b). Problematically, toxic geeks are often framed as harmless 'dorks' which obscures their complicity in the proliferation of hegemonic masculinity. Toxic geeks translate their inferiority to hegemonic men into attempts to demonstrate superiority over FLINTA* or queers which can result in sexist and queerphobic aggressions (McIntosh, 2017a,b). As Newton and Stanfill (2020) argue, toxic geek masculinity 'specifically employs technology as the way to approximate hegemonic masculinity'. This is particularly evident in the case of pornographic deepfakes, where DL technology becomes the tool to approximate sex with attractive women as it is scripted for hegemonic men.

In the manner of straightening devices, the software brings into reach what has been

[14] in terms of a traditional career

out of reach before: the intimacy with a desired person as well as the ability of subjects to conform with hegemonic gender scripts. The power, then, expresses itself not only in offering a tool to actualize hegemonic masculinity, but also in constructing a rational framework for justifying deepfake porn production that potently enters the discourse. One user in the reddit forum r/changemyview illustrates this by specifically using technology to legitimize their entitlement over women:

> [...] these things shouldn't be banned, let alone outlawed – they're just a natural progression of technology. In 20 years I'll be able to have virtual reality sex with your wife, or your teenage daughter, or YOU, or whatever, and you won't be able to do a damn thing about it, and that's as it should be. (taken from Winter and Salter, 2020)

To rephrase this statement in Ahmed's vocabulary, deepfake technology extends the bodily horizon of its users and allows to reach bodies that were out of reach before. When this space of actions is enforced through repetition, a normative line arises, which asks—or even requires—to be followed. The assumption of a linear progression of technology, as this user describes, can be understood as such a normative line. Actions that are 'in line' with the (envisioned) trajectory of technological development are consequently legitimized. This legitimization and even naturalization of misogyny is an essential strategy in order to consistently reconfirm and justify the dominance of men.

If pornographic deepfakes function as a tool to approximate hegemonic masculinity scripts it is little surprising that they are also used to assert one's status in the community. In an analysis of deepfake development discourses on Github and Reddit, Newton and Stanfill (2020) observe that inexperience in coding is expressed in subservient and apologetic ways while skillful programming is met with appreciation and praise. This equaling of performance with a person's worth can lead skilled programmers to arrogant exclamations like 'I don't need acknowledgement from bunch of noobs' [sic] (Newton and Stanfill, 2020).[15] Newton and Stanfill (2020) attribute this ambition to the cultural dominance of meritocracy, a narrative that promises equal opportunity based on individual striving, but in fact upholds social inequalities, as for instance Jo Littler (2013) convincingly argues. Meritocracy is fundamental to modern neo-liberalism and particularly pronounced all across technocultures, ranging from the Silicon valley to open source software development projects (Newton and Stanfill, 2020). It is likely that the current "hype" around AI, and especially the popularity of DL in the past years, also constituted influential discursive elements and contributed to the appeal of deepfakes. In this context of meritocratic technocultures and a hype surrounding DL, the ability to create high quality deepfakes

[15]The production of deepfakes is generally available for anyone. Up to this date, the source code is available on GitHub. Several deepfake softwares such as 'Faceswap' and 'FakeApp' include a convenient user interface and can be downloaded for free, making the production of deepfakes feasible even without coding experience. However, to obtain higher quality resolutions and more lifelike renderings a skillful fine-tuning of training hyper-parameters is an asset for better results.

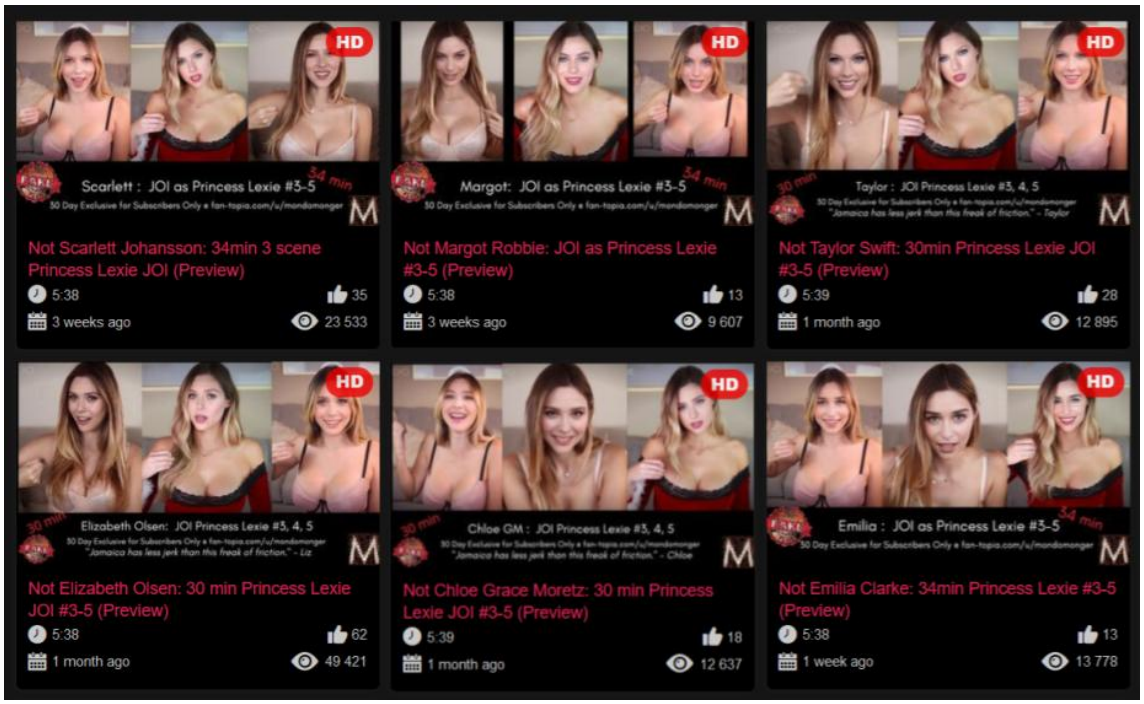

Figure 1: Results from deepfake porn platform mrdeepfakes.com. The same baseline video is uploaded several times featuring different celebrities.[17]

functions as a marker of technical versatility which offers prestige and status within the community. An interesting phenomenon on deepfake porn platforms is that creators often upload the same video multiple times, each one differing only in the celebrity that is featured (see Figure 1). This practice suggests that the incentive behind deepfake porn production is likely not the creation of an authentic sexual experience. Instead, different actresses are reduced to products which savvy creators use to showcase their expertise to the community.

These demonstrations of skill are a crucial component of another facet of deepfake porn. To thoroughly grasp the phenomenon of deepfake porn and its implications one must understand that deepfake porn is more than a lewd hobby for private amusement. On online platforms, it advances to a business model which re-arranges the material conditions of porn production. The importance of monetarization of deepfake porn is illustrated by the fact that numerous videos on one of the largest deepfake porn platforms contain the label 'paid request'[18]. This means that users can ask creators to produce a custom porn video of a celebrity, optionally also including a specific sexual practice or a porn performer who should lend their body to the celebrity's face. Users can come in contact with creators through online forums, the details are then usually exchanged in private messages.[19] From

[17] https://mrdeepfakes.com/search/pack Accessed on: 15.04.2021

[18] https://mrdeepfakes.com/search/paid-request Accessed on: 15.04.2021

[19] For curiosity, in my investigation I have contacted pornographic deepfake creators through such an online forum and inquired about the price of a custom deepfake porn. The prices varied greatly between 6$ and 50$ per minute of video.

the money producers earn with these videos they do not need to financially reimburse the people whose faces and bodies are depicted. Deepfake technology, thus, introduces a radical modification in the practice of porn production: it allows to continuously create novel content featuring femme bodies while at the same time completely circumventing these bodies from its payroll. Consequently, deepfake porn contributes to precarious labor conditions of porn performers and introduces yet another tool for exploiting femme bodies for someone else's material benefit.

The technology behind deepfakes has already advanced to a market offering various products (Jacoby, 09.12.2019). One instance is the app 'DeepNude': Instead of manipulating videos, the app removes clothing from photos, creating elusively realistic nude images. The app is designed in a way to only work on femme-presenting bodies (Cole, 2019). Unlike deepfakes, where a few hundred images (for instance the still frames of a short video) are required to abstract the facial features of the person, DeepNude only requires a single photo, making this form of image based gendered violence even more accessible, and hence dangerous (Cole, 2019). In an interview, the creator of DeepNude stated that he specifically developed the app with the aim to gain an 'economic return' after having had financial problems (Cole, 2019).

To sum up, pornographic deepfakes enforce hegemonic norms in various ways. Firstly, they reinforce the notion that consent and bodily autonomy of FLINTA* are dispensable and that their bodies can be used for others' enjoyment. Secondly, they contribute to a climate of violence which can be internalized by FLINTA* and cause them to align their behaviour with patriarchal expectations, reproducing norms of a submissive femininity. Thirdly, by building upon meritocratic and technocratic narratives ML-aided pornography offers a justification for the (ab)use of others as objects of one's own pleasure. They further enforce hegemonic conceptions of masculinity, allowing creators and consumers to approximate hegemonic masculinity and to demonstrate their status among peers through the marker of technical versatility. Lastly, they manifest the material exploitation of femme bodies and undermine fair compensation for porn performers.

5 Platforms

After having presented the severe effects of content moderation systems and pornographic deepfakes on FLINTA* and LGBTQ* communities, I turn to the role of platforms: How is the structure of social media platforms relevant to the propagation of hegemonic sexual norms? And how is power distributed across them to achieve these effects? Centrally, platforms as *intermediaries* are the sites where the various elements of content moderation and deepfake systems meet and intersect. As large-scale, heterogeneous apparatuses, these systems consist of users, an administration, user-generated content, moderation policies, externally hired moderators, ML models, as well as advertisers, legislation, and economic

incentives. Inherent characteristics of online platforms further distinctively shape the working of both content moderation and pornographic deepfakes, as I demonstrate in the following.

Content moderation is at the heart of online social media and networking platforms. As Tarleton Gillespie formulates it, content moderation, ultimately, 'is the essence of platforms, it is the commodity they offer' (Gillespie, 2018, p. 208). Without content moderation, platforms would aggravate repelling content and cease to be attractive to its users, hence lose their value as a platform. Likewise, it is precisely the participatory form of sharing content generated by users—which distinguishes platforms from other editorial media outlets—that creates the demand for moderation.

The phenomenon of deepfake porn is also closely tied to online platforms. First and foremost, commercial porn platforms are the sites where deepfake porn is circulated. Hence, the coverage of platforms and their ability to reach large audiences is what intensifies the harmful effects for the victims of deepfake porn. Additionally, online forums and collaboration platforms were also fundamental components in driving the development of deepfake technology. For instance, still up to this date, GitHub is hosting the source code for deepfake software, making this potently harmful software publicly available for anyone[20]. Online forums like Reddit and GitHub also aid users with the development of their own deepfakes, providing instructions and support from the community for the usage of this software (Winter and Salter, 2020). Further, the anonymity offered on platforms and the connectivity between different platforms impede the moderation of deepfake porn and make it hard to hold its creators accountable for the abuse. When banned from one site, deepfake videos, code and its communities can easily move to another one, making it particularly hard for those affected to counteract and have the content removed (Winter and Salter, 2020).

Platforms themselves are also entangled in dependencies on other platforms and intermediaries. Apple's App Store is one example for the far-reaching impact of an intermediary's *gatekeeping power* (Khan, 2018). To be admitted on iOS phones, apps need to pass through a review phase that is intentionally designed as a bottleneck (Gillespie, 2018, p. 84). The Apple management that erects the guidelines for the App Store is notorious for its blatant anti-sex stance and hinders any sexually explicit apps from entering (Gillespie, 2018, p. 79).[21] Tumblr had faced a similar fate: One month before the porn ban, Tumblr had been temporarily banned from the App Store due to allegations of hosting child exploitation material (Bronstein, 2020). Instead of merely removing the illegal child exploitation content, Tumblr banned sexually explicit content altogether, albeit legal and

[20]https://github.com/deepfakes/faceswap Accessed on: 12.04.2021

[21]In February 2010 alone, more than five thousand sexually suggestive apps were removed from the App Store (Gillespie, 2018, p. 84). Former CEO of Apple, Steve Jobs, makes his view plain in a response email to a user concerned about the company's involvement in moral policing: 'Folks who want porn can buy [an] Android phone' (Bosker, 20.06.2010).

vehemently supported by its users. After announcing the ban on pornography, Tumblr was re-admitted into the App Store (Bronstein, 2020). For Bronstein (2020) this illustrates 'Apple's significant influence over the content featured on a wide range of platforms, and the power it wields to restrict what people can easily find and view through its App Store policies'.

Additionally, online commercial social media platforms are embedded into a larger economic system which inflicts certain imperatives on them. In order to remain profitable, commercial platforms not only have to keep their user base and advertisers pleased, but also need to continuously attract new users and advertisers (Gillespie, 2018, pp. 18–20). In the case of Tumblr, many researchers have argued that the decision to remove pornography from the platform was influenced by advertisers' discomfort with the explicit materials and part of a plan to align Tumblr into a corporate advertising strategy after the company had been bought by Verizon (Bronstein, 2020; Southerton et al., 2020; Ashley, 2019; Pettis, 2020). Advertisers might be inclined toward the dominant sexual norm due to their socialization or strategically design their ads according to mainstream conceptions to appeal to larger crowds. In either case, economic constraints imposed on for-profit platforms can contribute to a reproduction and propagation of an existent (anti-porn) norm. Similarly, the popularity of deepfake porn can appear as a viable method for porn platforms to attract viewers.

The network effect, 'the phenomenon whereby a product or service becomes more valuable the more that users use it' (Khan, 2018) can lead a small number of platforms to advance to massive central players. This also affects the development and application of AI: the resources to develop models, gather sufficiently large data sets, and hire skilled ML engineers to operate them are harder to procure for smaller platforms, eventually privileging the major players (Cambridge Consultants, 2019). The network effect consolidates and improves the position of large established platforms, making it harder for outsiders to compete. As can be seen in the example of Apple, increasing market dominance also augments a platform's ability to exert gatekeeping power. By being able to control content and dictate a norm, large platforms advance to a status resembling Gramsci's description of 'cultural hegemony' in which dominant players utilize media's large audiences to disperse their ideas and to impose 'a general direction [...] on social life' (Gramsci, 2011, p. 145).

Platforms, its users, administration and advertisers are further inseparable from public discourses. Discourses around pornography shape platforms' content and direct their administrative direction. Likewise, platforms are spaces where pornographic discourses can be fostered, refined or reconfigured. Pornographic discourses might be as old as humans themselves, and are inextricably linked to the prevalent notions of sexuality. Genealogically tracing their influence on current trends in automated pornography production and censorship would thus by far exceed the scope of this work. Therefore, I will set aside

a thorough analysis of how public discourses shape platforms' stance toward pornography. Likewise, examining how the discursive spaces of platforms contribute to forming or altering a certain conception of pornography would also demand a separate work.

Yet, to illustrate the involvement of discourses on apparatuses, I briefly touch upon a common porn narrative to that may appear as a universal truth, but requires a critical examination of its strategic function and relation to the apparatus of content moderation. In the interview *Confessions of the Flesh*, Foucault describes how discourses contribute to the functioning of a strategy, and thus the operation of an apparatus. He observes how the seemingly positive narratives of 'philanthropy and the moralisation of the working class' (Foucault, 2008, p. 203) in 18th century France were crucially involved in mobilizing an otherwise reluctant labour force for the mercantilist state.[22] Similarly, narratives applied by platform administrations aimed at legitimizing moderation of explicit content allude to 'a better, more positive [platform]' (Tumblr Staff, 03.12.2018) that protects children (Bosker, 20.06.2010) and is 'suitable for all audiences' (blog.youtube, 20.03.2017). Ironically, these narratives often obscure how queer and other niche sex-positive/kink communities tend to be marginalized and stigmatized on the "big stage" for the mainstream audience (Southerton et al., 2020). They are further grounded in the assumption that pornography is an irritating and spoiling evil—a conception that must be questioned in its general validity and considered from the historical context of a larger strategy which aims to regulate sexuality, and eventually subjects (Foucault, 1978). This demonizing narrative is also the tenor in papers introducing automated pornography detection algorithms, which sketch pornography as threatening, and conclude that their technical solutions are urgently required to alleviate the hazard (e.g. Tang et al., 2009; Xu et al., 2020). This strategy can also be related to the argument of protecting children: Interestingly, the argument of protecting children from vulgar content is similarly adopted by conservatives to justify queerphobic legislation (Buyantueva, 2018). Queerphobic discourses often overlap with anti-porn discourses and are hard to disentangle from one another. Such narratives must thus be critically examined in their strategic function: As Vorhölter (2017) observes, laws and regulations aimed at protecting minors from sexuality can in fact operate as tools to regulate their sexuality and '[provide] [...] authorities with leverage to reinforce certain orders'. For both queerphobic and anti-porn stances, it can thus be hypothesized that claims to protect kids are part of a strategy to control a divergence of their interest, in order to guarantee that only desired forms of sexuality are reproduced.

Similarly, pornographic deepfakes owe a large part of their harm to dominant sex-negative discourses. While unwantedly appearing in a porn movie does harm people in their autonomy, the most sinister effects draw from social shaming that depicted people face for appearing in sexual contexts. This shaming and harassment is fueled by the

[22]This discourse was materialised in the erection of institutions to further propagate the values and norms required for restructuring the society (Foucault, 2008, p. 203).

discursive conception of sex as abnormal and immoral, as it is reproduced through content moderation on most social media platforms.

Although being far from an extensive analysis of discourses shaping content moderation and deepfakes, these instances indicate that discourses unmistakably shape the workings of the apparatuses I investigate.

To conclude, the ML systems of content moderation and pornographic deepfakes are inherently entangled with various online platforms. These platforms themselves are interwoven in dependencies on other platforms, which may dictate their policies onto them. Further, economic imperatives that demand an appeal to advertisers and principles like the network effect configure the working of platforms, and thus of content moderation and pornographic deepfakes. These systems are also crucially *shaped by* and *shaping* discourses, which can become part of the apparatuses and contribute to their strategy. This illustrates that none of the elements involved in the ML systems can be held responsible alone for the reproduction of certain norms. The agency implicated in norm (re-)production thus cannot be ascribed to a single actor, but is distributed across the system. Identifying and understanding how these components of the apparatus enable, constrain or depend on each other is paramount not only to comprehensively grasp the operation of power in content moderation systems and pornographic deepfakes, but also fundamental in order to interfere with their current ability to reinforce injustices.

6 Conclusion

> *We haven't made a dent in removing porn from the internet, but we've managed to steal the place where women commonly engage with other sex positive individuals; where transgender individuals may feel most secure; and where inclusivity was part of the unwritten contract by all who shared this kink-friendly home.* (Clark, 2018)

This work investigated how ML systems are involved in reproducing traditional (hegemonic) norms. To do so, I have selected two ML applications that are used for the production and deletion of pornography, content moderation system and deepfakes. In order to thoroughly grasp the operation and effects of these example applications, I have considered them as *socio-technical systems* comprising technical as well as societal elements, such as data, platforms, users, developers, discourses, legislation and economic incentives. Inspired by the work of Foucault, I have applied an understanding of power as a productive and repressive force, which operates through the heterogeneous arrangements of *apparatuses*. In addition, I have incorporated Sara Ahmed's notion of normativity-producing *straightening devices* that makes intelligible how content moderation systems and pornographic deepfakes are implicated in constructing and enforcing normative alignments.

By examining concrete effects of the ML systems of content moderation systems and pornographic deepfakes I have found that they are involved in reinforcing heteronormativity, hegemonic conceptions of masculinity and femininity, and manifesting exploitative material conditions, which are components of what I have termed *hegemonic sexuality*. These effects are achieved through an operation of power that acts both productively and repressively, and uses material as well as discursive mechanisms. The systems of content moderation systems and pornographic deepfakes can thus be understood as apparatuses which are arranged in a specific way to promote hegemonic understandings of sexuality and fend off alternative, counter-hegemonic practices and conceptions. Through this (re-)production of the hegemonic form of sexuality they contribute to upholding an existent asymmetry of power, which can be identified as their *strategy*.

Various of the effects I have assessed are not novel, but were already ingrained into existing practices of pornography. ML, however, contributes to amplify existing hegemonic norms or gives them new ways to express themselves, like in the monetarization of deepfake porn. What is further special about these automated approaches is the scale and velocity at which they allow to reinforce dominant norms. Automated content moderation systems allow to detect content inconsistent with the norm on a scale unfeasible for human moderators, making it possible to scrape off sex workers or queer porn from entire platforms. The automation present in pornographic deepfakes makes the production of image based sexualized violence widely accessible without requiring extensive technical skills.

It remains important to examine further ML applications on their ability to form and enforce norms. What can nonetheless be concluded for ML systems based on the ones I have observed is the way power is exerted through them: As I have illustrated using Foucault's concept of the apparatus, ML applications cannot be envisioned as mere tools of power. Rather, ML and its applications are in themselves intricately entangled with power, and are both product as well as origin of power.

In the process of this work, especially in the part of pornographic deepfakes, I have become more aware of the difficulty of writing about hegemonic sexuality without perpetuating its underlying narratives. When FLINTA* depicted in deepfake porn are only portrayed as passive victims and men framed as aggressive perpetrators, it can 'freez[e] women into powerless positions of rapability' (Gunnarsson, 2018) and reinforce hegemonic gendered conceptions rather than challenge them. I have tried to the best of my knowledge to give an exhaustive account of the the phenomenon of deepfake porn and tried to avoid falling into such dominant narratives. However, it is not impossible that I have, at times, unknowingly and unwillingly reiterated an understanding that victimizes FLINTA* as passive and powerless. Here, it again comes to light how deeply dominant conceptions of gender and sexuality are inscribed into one's conceptions, and how a critical analysis can never be outside of the normative alignments it is trying to assess.

As an outlook, while some platforms have prohibited non-consensual porn like deepfakes, the risk they pose for FLINTA* still remains. Likewise, it can be expected that content moderation will not lose its relevance any time soon, but continues to mediate our online interactions and to (re-)produce a specific, normatively aligned, reality. Content moderation alone will not be able to solve intricate societal problems like misogyny, queer- and transphobia, as it only tries to keep its symptoms out of sight, but does not eradicate its roots. Becoming aware of these powerful workings of apparatuses that further misogyny, queer- and transphobia can certainly seem discouraging and even paralyzing. However, in such moments of discouragement, Foucault (1978, p. 95) can remind us that 'where there is power, there is resistance'. The question thus remains whether the ensembles of content moderation, deepfakes and platforms can be re-arranged to subvert hegemonic conceptions and used to realize a more feminist, inclusive and liberatory vision of the world. How can online spaces be built and defended that offer safety, representation and empowerment to FLINTA*? How will content moderation systems be applied across platforms to better protect people affected by online sexualized violence? And how could AI and ML applications be re-configured to support sexual self-determination?

7 List of Abbreviations

AI Artificial Intelligence

ML Machine Learning

DL Deep Learning

GAN Generative Adversarial Network

FLINTA* Female Lesbian Intersex Non-Binary Transgender Agender *

LGBTQ* Lesbian Gay Bisexual Transgender Queer *

References

Sara Ahmed. *Queer phenomenology: Orientations, objects, others*. Duke Univ. Press, Durham, NC, 2006. ISBN 978-0-8223-8807-4.

Henry Ajder. Deepfake Threat Intelligence: a statistics snapshot from June 2020. *Sensity*, 2020. Accessed on: 30.01.2021. https://sensity.ai/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/.

Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The State of Deepfakes: Landscape, Threats, and Impact, 2019.

Richmond Alake. Deep Learning: GoogLeNet Explained - Towards Data Science. *Towards Data Science*, 23.12.2020. Accessed on: 08.05.2021. https://towardsdatascience.com/deep-learning-googlenet-explained-de8861c82765.

Lux Alptraum. Deepfake Porn Harms Adult Performers, Too. *WIRED*, 15.01.2020. Accessed on: 20.04.2021. https://www.wired.com/story/deepfake-porn-harms-adult-performers-too/.

Vex Ashley. Tumblr porn eulogy. *Porn Studies*, 6(3):359–362, 2019. ISSN 2326-8743.

Rana Ayyub. I Was The Victim Of A Deepfake Porn Plot Intended To Silence Me. *HuffPost UK*, 2018, 21.11.2018. Accessed on: 30.01.2021. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316.

Rana Ayyub. In India, Journalists Face Slut-Shaming and Rape Threats: Opinion. *The New York Times*, 2018, 22.05.2018. ISSN 1553-8095. Accessed on: 27.01.2021. https://www.nytimes.com/2018/05/22/opinion/india-journalists-slut-shaming-rape.html.

Julia Bähr. Das Epizentrum der Remix-Kultur. *Frankfurter Allgemeine Zeitung*, 13.02.2015. Accessed on: 05.04.2021. https://www.faz.net/aktuell/feuilleton/medien/zehn-jahre-youtube-epizentrum-der-remix-kultur-13418285.html.

Karen Barad. *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Duke Univ. Press, Durham, 2007. ISBN 9780822388128.

Sam Biddle, Paulo Victor Riberiro, and Tatiana Dias. TikTok Told Moderators to Suppress Posts by "Ugly" People and the Poor to Attract New Users. *The Intercept*, 16.03.2020. Accessed on: 22.03.2021. https://theintercept.com/2020/03/16/tiktok-app-moderators-users-discrimination/.

blog.youtube. Restricted Mode: How it works and what we can do better, 20.03.2017. Accessed on: 12.03.2021. https://blog.youtube/news-and-events/restricted-mode-how-it-works-and-what.

Angelos Bollas. Masculinities on the Side: An Exploration of the Function of Homosexism in Maintaining Hegemonic Masculinities and Sexualities. *Sexuality & Culture*, pages 1–10, 2021. ISSN 1936-4822. doi: 10.1007/s12119-021-09848-3. https://link.springer.com/article/10.1007/s12119-021-09848-3.

Bianca Bosker. Steve Jobs: 'Folks Who Want Porn Can Buy An Android Phone'. *HuffPost*, 20.06.2010. Accessed on: 18.05.2021. https://www.huffpost.com/entry/steve-jobs-reiterates-fol_n_544045.

Carolyn Bronstein. Pornography, Trans Visibility, and the Demise of Tumblr. *TSQ: Transgender Studies Quarterly*, 7(2):240–254, 2020. ISSN 2328-9252.

Jason Brownlee. *Master Machine Learning Algorithms: discover how they work and implement them from scratch*. Machine Learning Mastery, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

Radzhana Buyantueva. LGBT Rights Activism and Homophobia in Russia. *Journal of Homosexuality*, 65(4):456–483, 2018. doi: 10.1080/00918369.2017.1320167.

Cambridge Consultants. Use of AI in Online Content Moderation, 2019.

Adrian Chen. The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. *WIRED*, 23.10.2014. Accessed on: 04.04.2021. https://www.wired.com/2014/10/content-moderation/.

Alexander Cho. *Sensuous Participation: Queer Youth of Color, Affect, and Social Media*. Dissertation, University of Texas, 2015.

Danielle Keats Citron. The National Security Challenge of Artificial Intelligence, Manipulated Media, and "Deep Fakes": Prepared Written Testimony and Statement for the Record, 2019.

Bryan Clark. Tumblr's porn ban slams the door on women and other marginalized communities, 2018. Accessed on: 24.05.2021. https://thenextweb.com/news/tumblrs-porn-ban-slams-the-door-on-marginalized-communities.

Jennifer Cobbe. Algorithmic Censorship by Social Platforms: Power and Resistance. *Philosophy & Technology*, pages 1–28, 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00429-0. https://link.springer.com/article/10.1007/s13347-020-00429-0.

Samantha Cole. Deepnude: The Horrifying App Undressing Women, 2019. Accessed on: 18.04.2021. https://www.vice.com/en/article/kzm59x/deepnude-app-creates-fake-nudes-of-any-woman.

R. W. Connell. Masculinities and Globalization. *Men and Masculinities*, 1(1):3–23, 1998. doi: 10.1177/1097184X98001001001.

Reyhan Şahin. *Yalla, Feminismus!* Tropen, 2019.

Cookie Cyboid. Want To Know Why Tumblr Is Cracking Down On Sex? Look To FOSTA/SESTA, 2018. Accessed on: 02.03.2021. https://medium.com/the-establishment/want-to-know-why-tumblr-is-cracking-down-on-sex-look-to-fosta-sesta-15c4174944a6.

Thiago Dias Oliva, Dennys Marcelo Antonialli, and Alessandra Gomes. Fighting Hate Speech, Silencing Drag Queens? Artificial Intelligence in Content Moderation and Risks to LGBTQ Voices Online. *Sexuality & Culture*, 25(2):700–732, 2021. ISSN 1936-4822. doi: 10.1007/s12119-020-09790-w.

Stephanie Dick. Artificial Intelligence. *Harvard Data Science Review*, 1(1), 2019. Accessed on: 07.05.2021. https://hdsr.mitpress.mit.edu/pub/0aytgrau/release/2?readingCollection=672db545.

Gail Dines, Bob Jensen, Robert Jensen, and Ann Russo. *Pornography: The production and consumption of inequality*. Psychology Press, 1998.

Mike Donaldson. What is hegemonic masculinity? *Theory and Society*, 22(5):643–657, 1993. ISSN 0304-2421. doi: 10.1007/bf00993540.

Suzie Dunn. *Technology-Facilitated Gender-Based Violence: An Overview*. 2020.

Andrea Dworkin. Against the male flood: Censorship, pornography, and equality. *Harv. Women's LJ*, 8, 1985.

Dossie Easton and Catherine A Liszt. *The ethical slut: A guide to infinite sexual possibilities*. Greenery Press San Francisco, CA, 1997.

Chappell Ellison. Twitter status, 07.12.2018. Accessed on: 11.03.2021. https://twitter.com/ChappellTracker/status/1071099239353249792.

Jacob Engelberg and Gary Needham. Purging the queer archive: Tumblr's counter-hegemonic pornographies. *Porn Studies*, 6(3):350–354, 2019. ISSN 2326-8743. doi: 10.1080/23268743.2019.1623067.

Facebook for Media. Moderate Your Facebook Page, 2021. Accessed on: 13.04.2021. https://www.facebook.com/formedia/blog/moderating-your-facebook-page.

Don Fallis. The Epistemic Threat of Deepfakes. *Philosophy & Technology*, pages 1–21, 2020. ISSN 2210-5441. doi: 10.1007/s13347-020-00419-2. https://link.springer.com/article/10.1007/s13347-020-00419-2.

Gretchen Faust. Hair, Blood and the Nipple. *Digital Environments*, pages 159–170, 2017. ISSN 97838394. https://www.degruyter.com/document/doi/10.14361/9783839434970-012/html.

Silvia Federici. *Caliban and the witch: [women, the body and primitive accumulation]*. Autonomedia, Brooklyn NY, 1. ed. edition, 2004. ISBN 1570270597.

Hinrich Fink-Eitel. *Foucault zur Einführung*. Ed. SOAK im Junius-Verlag, 1989.

Michel Foucault. *The history of sexuality, Vol. 1: An introduction*. 1978.

Michel Foucault. *Truth and Power: Interview with Alessandro Fontano and Pasquale Pasquino*. Feral Publications, Sydney, 1979.

Michel Foucault. *Order of Things*. Routledge Classics. Taylor & Francis Group, London, 2nd ed. edition, 2005. ISBN 9780203996645. doi: 10.4324/9780203996645. https://www.taylorfrancis.com/books/mono/10.4324/9780203996645/order-things-michel-foucault.

Michel Foucault. Power/Knowledge. In Jeffrey C. Alexander and Steven Seidman, editors, *The new social theory reader*, pages 73–79. Routledge, London and New York, 2008. ISBN 9781003060963. doi: 10.4324/9781003060963-10. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003060963-10/power-knowledge-michel-foucault.

Michel Foucault. *Discipline and punish: The birth of the prison*. Vintage, New York, 2012. ISBN 9780307819291.

Robert W. Gehl, Lucas Moyer-Horner, and Sara K. Yeo. Training Computers to See Internet Pornography: Gender and Sexual Discrimination in Computer Vision Science. *Television & New Media*, 18(6):529–547, 2017. doi: 10.1177/1527476416680453.

Tarleton Gillespie. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, New Haven, 2018. ISBN 9780300235029.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Antonio Gramsci. *Prison Notebooks*, volume 2. Columbia University Press, 2011.

Lena Gunnarsson. "Excuse Me, But Are You Raping Me Now?" Discourse and Experience in (the Grey Areas of) Sexual Violence. *NORA - Nordic Journal of Feminist and Gender Research*, 26(1):4–18, 2018. doi: 10.1080/08038740.2017.1395359.

Anita Gurumurthy, Amrita Vasudevan, and Nandini Chami. Born digital, Born free? A socio-legal study on young women's experiences of online violence in South India. Accessed on: 27.03.2021. `https://itforchange.net/sites/default/files/1662/Born-DigitalBorn-FreeSynthesisReport.pdf`.

bell hooks. Penis Passion. *Lion's Roar*, 01.07.1999. Accessed on: 19.04.2021. `https://www.lionsroar.com/penis-passion/`.

Monika Jäckle. Geschlechterdispositiv, 2014. Accessed on: 12.04.2021. `https://gender-glossar.de/g/item/35-geschlechterdispositiv`.

Evan Jacoby. I Paid $30 to Create a Deepfake Porn of Myself. *VICE*, 09.12.2019. Accessed on: 20.04.2021. `https://www.vice.com/en/article/vb55p8/i-paid-dollar30-to-create-a-deepfake-porn-of-myself`.

Sasan Karamizadeh and Abouzar Arabsorkhi. Methods of Pornography Detection. In *Proceedings of the 10th International Conference on Computer Modeling and Simulation*, ACM Other conferences, New York, NY, 2018. ACM. ISBN 9781450363396. doi: 10.1145/3177457.3177484.

Lina M. Khan. Sources of Tech Platform Power. *2 Georgetown Law Technology Review 325*, 2018. Accessed on: 25.01.2021.

Chris Köver. Discrimination - TikTok curbed reach for people with disabilities. 02.12.2019. Accessed on: 22.03.2021. `https://netzpolitik.org/2019/discrimination-tiktok-curbed-reach-for-people-with-disabilities/`.

Susan Leavy. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st international workshop on gender equality in software engineering*, pages 14–16, 2018.

Rachael Liberman. 'it's a really great tool': feminist pornography and the promotion of sexual subjectivity. *Porn Studies*, 2(2-3):174–191, 2015.

Jo Littler. *Meritocracy as Plutocracy: The Marketising of 'Equality' Under Neoliberalism*, volume 80. Lawrence and Wishart, 2013. doi: 10.3898/NewF.80/81.03.2013. https://www.ingentaconnect.com/content/lwish/nf/2013/00000080/00000080/art00005.

Emma Llansó, Joris van Hoboken, Paddy Leersen, and Jaron Harambam. Artificial intelligence, content moderation, and freedom of expression. *Transatlantic Working Group*, 2020.

Catherine A. MacKinnon. Sexuality, Pornography, and Method: Pleasure under Patriarchy. *Ethics*, 99(2):314–346, 1989. doi: 10.1086/293068.

Sophie Maddocks. 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, 7(4):415–423, 2020. ISSN 2326-8743.

Louise Matsakis. Tumblr's Porn-Detecting AI Has One Job—and It's Bad at It. *WIRED*, 06.12.2018. Accessed on: 02.06.2021. https://www.wired.com/story/tumblr-porn-ai-adult-content/.

Jonathan McIntosh. The Adorkable Misogyny of The Big Bang Theory, 2017a. Accessed on: 12.04.2021. http://popculturedetective.agency/2017/the-adorkable-misogyny-of-the-big-bang-theory.

Jonathan McIntosh. Complicit Geek Masculinity and The Big Bang Theory, 2017b. Accessed on: 12.04.2021. http://popculturedetective.agency/2017/complicit-geek-masculinity.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

Rainer Mühlhoff. *Immersive Macht: Affekttheorie nach Spinoza und Foucault*. Campus Verlag, Frankfurt am Main, 1. auflage edition, 2018. ISBN 9783593438597.

Rainer Mühlhoff. *Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning*. Technische Universität Berlin, Berlin, 2021. doi: 10.14279/depositonce-11329.

Neonfiona. Twitter status: Just looked at my videos with the restricted mode on. Seeing a bit of a theme here... LGBT+ content not safe for kids @YouTube?, 16.03.2017. Accessed on: 12.03.2021. https://twitter.com/neonfiona/status/842390135257874432.

Dorothy Neufeld. The 50 Most Visited Websites in the World. *Visual Capitalist*, 27.01.2021. Accessed on: 22.03.2021. https://www.visualcapitalist.com/the-50-most-visited-websites-in-the-world/.

Olivia B. Newton and Mel Stanfill. My NSFW video has partial occlusion: deepfakes and the technological production of non-consensual pornography. *Porn Studies*, 7(4): 398–414, 2020. ISSN 2326-8743. doi: 10.1080/23268743.2019.1675091.

Fudong Nian, Teng Li, Yan Wang, Mingliang Xu, and Jun Wu. Pornographic image detection utilizing deep convolutional neural networks. *Neurocomputing*, 210:283–293, 2016. ISSN 0925-2312. doi: 10.1016/j.neucom.2015.09.135. https://www.sciencedirect.com/science/article/pii/S0925231216305963.

Mirko Nikolić. *minoritarian ecologies: performance before a more-than-human world*. PhD thesis, 2017. https://westminsterresearch.westminster.ac.uk/item/9zyyy/minoritarian-ecologies-performance-before-a-more-than-human-world.

Johanna Oksala. *Foucault on freedom*. Modern European philosophy. Cambridge University Press, Cambridge, 2005. ISBN 9780521847797. http://www.loc.gov/catdir/enhancements/fy0642/2005047062-d.html.

Omnicore. Facebook by the Numbers: Stats, Demographics & Fun Facts, 2021. Accessed on: 14.04.2021. https://www.omnicoreagency.com/facebook-statistics/.

Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Allen Lane, UK and USA, first edition edition, 2016. ISBN 9780241296813.

Susanna Paasonen. *Carnal resonance: Affect and online pornography*. MIT Press, Cambridge, Mass and London, 2011. ISBN 1283321769. http://lib.myilibrary.com/detail.asp?id=332176.

Ben Pettis. *Tumblr Porn Ban: the platform triad and the shaping of online spaces, The*. PhD thesis, Colorado State University, 2020.

Marianne Pieper and Robin Bauer. Polyamorie: Mono-normativität–dissidente mikropolitik–begehren als transformative kraft? *Journal für Psychologie*, 22(1), 2014.

Elena Pilipets and Susanna Paasonen. Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society*, page 146144482097928, 2020. ISSN 1461-4448. doi: 10.1177/1461444820979280.

Pinscreen. AI-Driven Virtual Avatars, 2020. Accessed on: 01.04.2021. https://www.pinscreen.com/facereplacement/.

Pornhub. Reporting Abuse And Violations, 03.04.2021. Accessed on: 03.04.2021. https://help.pornhub.com/hc/en-us/sections/360006831174-Reporting-Abuse-And-Violations.

Pornhub Insights. The 2019 Year in Review, 2019. Accessed on: 14.04.2021. https://www.pornhub.com/insights/2019-year-in-review.

Annie Potts. "The Essence of the Hard On". *Men and Masculinities*, 3(1):85–103, 2000. doi: 10.1177/1097184X00003001004.

William J Rapaport. Semiotic systems, computers, and the mind: How cognition could be computing. *International Journal of Signs and Semiotic Systems (IJSSS)*, 2(1):32–71, 2012.

Sarah T. Roberts. Digital detritus: 'Error' and the logic of opacity in social media content moderation, 2018. Accessed on: 04.02.2021. `https://journals.uic.edu/ojs/index.php/fm/article/download/8283/6649#2a`.

Laurie Shrage. Exposing the fallacies of anti-porn feminism. *Feminist Theory*, 6(1):45–65, 2005.

SimilarWeb. Top Websites - Website Ranking by Traffic — SimilarWeb, 22.03.2021. Accessed on: 22.03.2021. `https://www.similarweb.com/top-websites/`.

Spandana Singh. Everything in Moderation, 2019. Accessed on: 18.12.2020. `https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/introduction/`.

Ingrid Solano, Nicholas R. Eaton, and K. Daniel O'Leary. Pornography Consumption, Modality and Function in a Large Internet Sample. *Journal of sex research*, 57(1): 92–103, 2020. doi: 10.1080/00224499.2018.1532488.

Clare Southerton, Daniel Marshall, Peter Aggleton, Mary Lou Rasmussen, and Rob Cover. Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society*, page 146144482090436, 2020. ISSN 1461-4448.

Tamsin Spargo. *Foucault and queer theory*. Postmodern encounters. Icon Books and Totem Books, Cambridge and New York NY, 1999. ISBN 184046092X.

Sheng Tang, Jintao Li, Yongdong Zhang, Cheng Xie, Ming Li, Yizhi Liu, and Xiufeng Hua. Pornprobe: an lda-svm based pornography detection system. 2009.

Tristan Taormino, Constance Penley, Celine Shimizu, and Mireille Miller-Young. *The feminist porn book: The politics of producing pleasure*. The Feminist Press at CUNY, 2013.

The Recovery Village. Pornography Facts and Statistics, 2021. Accessed on: 29.03.2021. `https://www.therecoveryvillage.com/process-addiction/porn-addiction/related/pornography-statistics/`.

Heidi Tripp. All Sex Workers Deserve Protection: How FOSTA/SESTA Overlooks Consensual Sex Workers in an Attempt to Protect Sex Trafficking Victims. *Penn State Law*

Review, 124:219, 2019. https://heinonline.org/HOL/Page?handle=hein.journals/dlr124&id=227&div=9&collection=journals.

Tumblr. Adult content, 09.01.2021. Accessed on: 09.01.2021. https://tumblr.zendesk.com/hc/en-us/articles/231885248-Adult-content.

Tumblr. Community Guidelines, 16.07.2020. Accessed on: 09.01.2021. https://www.tumblr.com/policy/en/community.

Tumblr Staff. A better, more positive Tumblr, 03.12.2018. Accessed on: 03.06.2021. https://staff.tumblr.com/post/180758987165/a-better-more-positive-tumblr.

Tumblr Support. Updates to Tumblr's Community Guidelines, 2018. Accessed on: 09.01.2021. https://support.tumblr.com/post/180758979032/updates-to-tumblrs-community-guidelines.

Franz von Kutschera. *Einführung in die Logik der Normen, Werte und Entscheidungen*. 1973. https://epub.uni-regensburg.de/12511/1/ubr05427_ocr.pdf.

Julia Vorhölter. Homosexuality, pornography, and other 'modern threats' – The deployment of sexuality in recent laws and public discourses in Uganda. *Critique of Anthropology*, 37(1):93–111, 2017. doi: 10.1177/0308275X16682601.

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018:7068349, 2018. ISSN 1687-5265. doi: 10.1155/2018/7068349. https://www.hindawi.com/journals/cin/2018/7068349/.

Travis L. Wagner and Ashley Blewer. "The Word Real Is No Longer Real": Deepfakes, Gender, and the Challenges of AI-Altered Video. *Open Information Science*, 3(1):32–46, 2019. ISSN 2451-1781. doi: 10.1515/opis-2019-0003. https://www.degruyter.com/document/doi/10.1515/opis-2019-0003/html.

Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2):246, 2018.

Rachel Winter and Anastasia Salter. DeepFakes: uncovering hardcore open source on GitHub. *Porn Studies*, 7(4):382–397, 2020. ISSN 2326-8743. doi: 10.1080/23268743.2019.1642794.

Wei Xu, Hamid Parvin, and Hadi Izadparast. Deep Learning Neural Network for Unconventional Images Classification. *Neural Processing Letters*, 52(1):169–185, 2020. ISSN 1573-773X. doi: 10.1007/s11063-020-10238-3. https://link.springer.com/article/10.1007/s11063-020-10238-3.