Algorithmic Censorship of Art: A Proposed Research Agenda

Piera Riccio
ELLIS Alicante
Alicante, Spain
piera@ellisalicante.org

Jose Luis Oliver
Architecture Department
Universidad de Alicante
Alicante, Spain
joseluis.oliver@ua.es

Francisco Escolano
Computer Science Department
Universidad de Alicante
Alicante, Spain
sco@ua.es

Nuria Oliver
ELLIS Alicante
Alicante, Spain
nuria@ellisalicante.org

Abstract

In the past decade, the application of Artificial Intelligence (AI) techniques to autonomously generate creative content or to support human creativity has gained interest from the scientific community. The generative models that have been proposed in the literature are changing the agency and dynamics of our art practices. A less explored area in the intersection of AI and creativity includes the indirect impact of AI on our creativity through content moderation algorithms on social media. Such algorithms tend to censor artistic pieces that display nudity, acting as inhibitors of human creativity. In this paper, we present a research agenda to tackle this challenge from a cultural and gender perspective, and we propose that a human and humanities-centered approach is necessary to develop AI systems that positively impact artistic practices.

Introduction

Social media platform adoption has grown exponentially in the past decade. Today, it is estimated that over 4.6 billion people in the world are active social media users[1]. For many of their users, these platforms have become the main source not only of social interactions, information and news (Walker and Matsa 2021), but also of their creative production and exposure to artistic content.

Artificial Intelligence (AI)-based algorithms are pervasive in social media platforms, to e.g. provide a personalized experience to their users, enable content search, target advertisements or automatically edit/filter images and videos. Content moderation[2] algorithms are a prominent example (Chen 2021). Protecting online users –particularly minors– from damaging content (e.g. violence, terrorism, hatred or pornography) is essential. Thus, most social media platforms publish community guidelines that define their content moderation policies. However, the immense volume of content posted and consumed daily on these platforms (e.g. over 90 million photos are posted on Instagram every day and more than 1 billion videos are viewed on TikTok daily) have led social media companies to heavily rely on AI-based algorithms for content moderation. Beyond inappropriate content, these algorithms tend to censor artistic pieces that display nudity –even when their intent is clearly non-sexual– constraining not only the freedom of expression of artists but also the cultural experiences of users.

Social media censorship concerns several aspects of our society and it is applied on a variety of artistic expressions. However, in this debate paper, we focus solely on the censorship of artistic nudity and we hypothesize that such censorship has a negative impact on the creative freedom of artists and on the broad diffusion of artistic content, eventually harming the users that they are trying to protect. As an example of such an impact, the Vienna museums created in 2021 an account on OnlyFans, an adult-only platform, after seeing their most famous artworks (by known artists, such as Schiele, Munch or Modigliani) repeatedly banned on Instagram, TikTok, and Facebook (Hunt 2021). The boundary between artistic nudes and pornography is highly debated among art theorists and sociologists (Vasilaki 2010; Patridge 2013; Eck 2001) and such an ambiguity is at the base of the cultural issue that we are addressing in our research.

In addition to the impact on the users of the platforms, several authors in the Computational Creativity (CC) community have argued that creativity needs to be situated and embodied in specific conditions to flourish (Saunders and Bown 2015; Guckelsberger et al. 2021). Considering social networks as a possible example of such an embodiment, censorship can have an impact on the inspiration for creative work not only for human authors but also for autonomous or co-creative systems that are immersed in this virtual environment, changing the nature of the artefacts that the system would be exposed to (Ritchie 2007). This negative impact of AI algorithms on social media contrasts the efforts of the scientific community, which in the past decade has shown great interest towards the development of AI algorithms that automatically generate art or assist humans in their creative processes. However, there is yet limited work in understanding the impact that such AI algorithms have on the cultural identity of our society. We believe that this subject deserves more attention from the computational creativity community. Hence, this short debate paper.

[1] https://datareportal.com/reports/digital-2022-global-overview-report

[2] Content moderation refers to the automatic prioritization, filtering, shadow-banning or censuring of content by means of AI-based algorithms.

Related Work

Social media is redefining the art world, from the marketing to the creation and curation of art. While these new dynamics and the democratization of art could be positive (Polaine, Street, and Paddington 2005), some authors claim that social media platforms have a negative impact on artistic production (James 2014) and creativity (Sharlow 2015). Manovich provides an overview of the connection between AI algorithms and the cultural ecosystems, emphasizing that the pervasiveness of AI algorithms is shaping our aesthetic decisions in creative media (Manovich 2018).

The algorithmic censorship of nudity on social media has been studied by several scholars, who have highlighted the disproportionate impact of such censorship on feminist artists (Faust 2017), and have explored the adopted artistic techniques to circumvent it (Olszanowski 2014). In recent years, artistic movements have emerged to publicly denounce the issue, such as *Don't Delete Art*[3] and *Artists Against Censorship*[4]. These initiatives and research related to this topic are of crucial importance to raise public awareness and to highlight the anthropological and sociological consequences of artistic censorship in social media. However, to the best of our knowledge, none of the existing initiatives address algorithmic censorship of art from a multidisciplinary perspective, including a technical analysis of the functioning of the content moderation algorithms.

AI-based algorithmic content moderation poses several societal challenges: first, such proprietary, machine learning-based algorithms are developed and maintained by private companies with clear economic incentives. Thus, their unprecedented power on defining our culture is exercised without any guarantee that it reflects the interests of society at large (Elkin-Koren 2020). Second, the automated decisions made by such algorithms are not always explainable and transparent, particularly if based on deep learning models. Third, algorithms are not foolproof and might not only make mistakes but also be fooled (Elkin-Koren 2020). Fourth, while historically controversial artistic content could be publicly discussed and debated, today artists have a limited ability to respond to censorship by social media platforms. Given the lack of transparency, it is hard to engage in a public debate if the reasons why certain content is banned are unknown. In contrast to related work, we propose a comprehensive research agenda on algorithmic censorship of art. Our objectives include an in-depth analysis of exemplary censored content, and the design of socio-technical solutions to mitigate such censorship.

AI and Art Censorship: A Historic Perspective

Nudity in the arts is historically considered *one of the defining aspects of mankind's creativity* (Deprez 2020). However, artistic nudes have been perceived, appreciated and accepted differently throughout history. Ancient Greeks conceived nudity as an expression of inner excellence, elevating humans from the realm of the flesh to the realm of Gods. In the Middle Ages, the same representations were perceived as obscene and sinful. In this period, classic paintings were covered and statues mutilated (Deprez 2019).

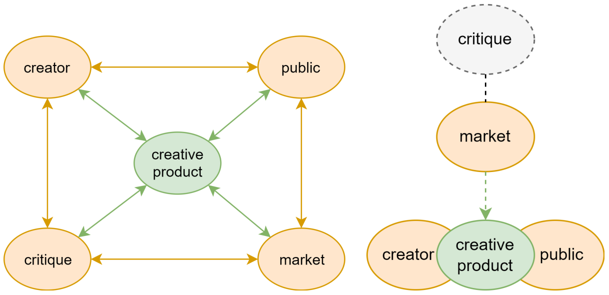

Figure 1: Synthetic sketch of the key elements within the creative ecosystem. Left, non-hierarchical arrangement among these elements before the advent of social media and AI; Right, transformation of the relationships in the context of AI algorithms used on social media.

The two aforementioned examples suggest that an understanding of the cultural context and ideals is necessary to embrace and appreciate the value of an artistic nude. Such context generally involves four key elements (critique/theory/context, market, public/observer and creators) to yield the creative product, as depicted in Figure 1. Historically (Figure 1, Left), these elements have been organized in a non-hierarchical structure, with connections among them. Depending on the artistic movement and the historic moment, one of these elements (for example the critique/theory) might have been more prominent that the rest in defining the environment for creativity (Montaner 1999). Studies in history of art identify and define the links and relations (depicted as arrows in the Figure) between the elements, and articulate a discourse about the artistic production from the perspective of different disciplines, including philosophy, morality, religion, politics, economics and aesthetics. Identifying the key elements and their relationships is crucial to develop a critical viewpoint of each creative framework, and to propose alternatives to it (Ramirez 1998). Today, these elements play new roles: the *public* is not simply a consumer, but it may become the product, i.e. the creation. Moreover, AI algorithms do not simply act as the *creators* (generating artistic content) but they can be, at the same time, the *critics* (deciding what is acceptable, and what is not) in a non-transparent way. We hypothesize that the ubiquity of opaque AI algorithms that impact the roles and links between the essential elements of the artistic creation environment hinders human creativity.

Art history is rich in examples of creative practices arisen from transgression and provocation towards existing ideals of morality. One such example is Michelangelo: despite working at the service of the Papacy, he depicted several nude figures in the iconic Sistine Chapel placing his masterpiece at risk of destruction (Vasari 1550). Unfortunately, disruptive artistic content might become an increasingly rarer phenomenon in our contemporary cultural environment (depicted in Figure 1, Right). AI algorithms, in fact, have the potential to not only influence one link in the diagram of the Figure 1, but simultaneously impact all the el-

[3] https://dontdelete.art/

[4] https://www.artistsagainstcensorship.com/

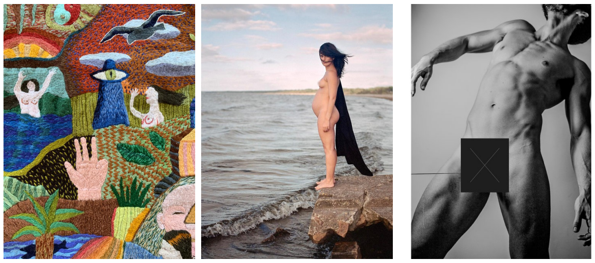

Figure 2: Three examples of censored images. Authors from left to right: Caroline Krabbe (collected through our survey), Adey (available on the *Artists Against Censorship* website), Udaentro (available on the *Don't Delete Art* website).

ements in the creative environment (Kulesz 2018). As a consequence, the traditional non-hierarchical structure morphs into a hierarchical organization where the *Market* lies on the top of the hierarchy, as the ultimate driver of the process, and therefore, as a fundamental agent in the creative decision-making process. Social media platforms are establishing a sort of monopoly to share content to the public. Algorithmic moderation on such platforms suffers from several important limitations: among others, we hypothesize that the utilized algorithms are unable to appreciate the value of an artwork or to understand the intent and context in which it is realized. As a consequence, social media leave no space for what is *blurred* (Kosko 1999) or *faint* (Vattimo 1988), drawing more defined –and yet invisible– lines between the acceptable and the unacceptable. In such a binary environment, breaking the rules is becoming harder, if not impossible.

AI and Art Censorship: Research Agenda

Given the importance of nudity in our artistic expression, we propose a research agenda on the topic of AI and algorithmic art censorship, articulated around four research questions.

RQ1: Pervasiveness of algorithmic censorship on social media

The first research question focuses on the pervasiveness of artistic nudity censorship on social medial platforms, its scope and characteristics.

Quantitative research in this domain is limited by the lack of representative, publicly available data, due to the proprietary nature of the social platforms and their content moderation algorithms. Hence, the first step in our research agenda entails reaching out to artist communities to collect a large corpus of censored artworks from social media. We are both establishing collaborations with relevant artists who have experienced censorship of their work and collecting additional examples of censored art through an online survey[5], which we launched in March of 2022.

The goal of this collection is to have a solid basis to shed light on the functioning of the content moderation algorithms and provide valuable feedback to artists as to why their content might have been shadow-banned or censored.

Preliminary analyses on the artworks that we have gathered to date reveal examples that depict female nudity with naivete (see first example in Figure 2), nudity without any sexual intent (see second example in Figure 2), or nudity that is already censored by the artist (see third example Figure 2). These pictures illustrate the extent of the issue that we plan to computationally analyze through the dataset.

RQ2: Human vs algorithmic censorship

The second research question aims to investigate the differences between the moral ideals embedded in today's content moderation algorithms and the human perception of art.

In 2021, the Facebook papers provided evidence that Meta maintains a *white list* of users[6] for which such content moderation rules do not apply. The inclusion in such a white list depends on the number of followers and popularity of a particular user. To highlight the market-driven decisions-making processes of content moderation algorithms, we plan to design and deploy a user study to collect ground truth on the appropriateness of the censored images (included in the dataset previously collected) when compared to other non-censored images displaying nudity. This research question aims to highlight the ability of people to recognize artistic intent in art and to show the existence of double standards on social media platforms. Given the broad reach of social media platforms across the planet, this user study will include a diverse set of participants from different cultural contexts to reflect the diversity of users in the platforms.

RQ3: Improved content moderation algorithms

Once we have a deeper understanding of the challenge, we plan to develop intent and context-aware content moderation algorithms that are able to distinguish artistic nudes from pornography.

Note that most of the social media platforms today do not explicitly ban artistic nudity in their community guidelines[7]. The discrepancy between the intent of the platforms and the actual censorship suggests that these algorithms are not yet refined enough to replace human moderators. In this regard, there is a need to develop content moderation algorithms that are intent and context-aware, combining different modalities (e.g. images and text) and leveraging inferred insights from the user study developed to address RQ2. Unfortunately, the existing ambiguity between artistic nudes and pornography is usually not taken into account by researchers developing algorithms for adult-content recognition (Wang et al. 2018; Chen 2021). We argue that an exploration of this issue could offer an opportunity in the field of Computational Creativity. In particular, the development of better content moderation algorithms for artistic nudity could leverage and improve the internal processes of evaluation in CC systems (Ventura 2017).

RQ4: Gender perspective

With RQ4, we address this topic with a gender perspective. The focus here is on studying the impact of such algorithms

[5] https://ellisalicante.org/censorship

[6] https://www.wsj.com/podcasts/the-journal Episode 1

[7] https://transparency.fb.com/it-it/policies/community-standards/adult-nudity-sexual-activity

on the cultural identity of women.

Throughout human history, women have been objectified in visual creative expressions (Barolsky 1999). While this pattern reappears with varied connotations in different historic time periods, the broad use of AI-based algorithms on social media could have unprecedented negative consequences for women. Remarkable feminist movements –such as *Free the Nipple* and *The Guerrilla Girls* (Pollen 2021)– have tried to raise social awareness about this issue.

In 1975, Mulvey (Mulvey 1975) identified the so called *male gaze* in Hollywood movies. This concept refers to a masculine heterosexual perception of women, who are depicted as objects of sexual desire, to satisfy what is known as *scopophilia* (i.e. the pleasure in looking). The concept of *male gaze* is still debated in today's visual culture. With the rise of social media, the *male gaze* has been argued to be stronger than it has ever been (Oliver 2017). We hypothesize that the censorship of female artistic nudes (and nipples, in particular) by AI algorithms has a role in this phenomenon. Today's AI algorithms on social media may be seen as socio-technical phenomena that automate culture through technology, perpetrating and possibly even amplifying human biases (Sezen 2020; Schroeder 2021). In particular, the censorship of female artistic nudity may be related to the conception of women as objects of pleasure. Because of this conception, female manifestations of nudity are frequently perceived as pornographic acts (Volkers 2020; Ibrahim 2017; Are 2021). This bias affects the freedom of expression of artists who are not conforming with the *male gaze* and that use female nudity to stand against the patriarchal sexualization of feminine bodies. Thus, we believe that the intersection between AI, social media, female nudity and art deserves to be further studied with a multi-disciplinary approach and a gender perspective.

Conclusion

In this paper we advocate for a research agenda focusing on the interplay between AI-based content moderation algorithms and art censorship on social media, and its implications on artistic production, creativity and the cultural identity of women. We have identified four broad research questions that would need to be addressed to fully understand such an interplay. These research questions (for example, the importance of having intent and context-aware content moderation algorithms) would need to be tackled *before* the widespread deployment of these technologies. Such a prior analysis would also entail interdisciplinary teams with experts from a variety of fields within the humanities and computer science (Crossick 2020). We emphasize the need to broaden the views of this research field, including both computing and non-computing disciplines (e.g. sociology, media studies, art history, anthropology) in the research agenda to develop technical solutions that are socially acceptable and responsible.

Author Contributions

P.R. and N.O. contributed to the proposal of the framework, proposal of the research questions, writing of the paper and its revision. J.L.O. contributed to the elements regarding the historic perspective. F.E. contributed to the paper's revision.

Acknowledgments

N.O. and P.R. are supported by a nominal grant received at the ELLIS Unit Alicante Foundation from the Regional Government of Valencia in Spain (Convenio Singular signed with Generalitat Valenciana, Conselleria d'Innovació, Universitats, Ciència i Societat Digital, Dirección General para el Avance de la Sociedad Digital). P.R. is also supported by a grant by the Banc Sabadell Foundation.

References

Are, C. 2021. The shadowban cycle: an autoethnography of pole dancing, nudity and censorship on instagram. *Feminist Media Studies* 1–18.

Barolsky, P. 1999. Looking at venus: A brief history of erotic art. *Arion: A Journal of Humanities and the Classics, (7):93–117.*

Chen, T. M. 2021. Automated content classification in social media platforms. In *Securing Social Networks in Cyberspace*. CRC Press. 53–71.

Crossick, G. 2020. From bridges to building sites: facilitating interdisciplinarity in the arts & humanities, last access: 19 may 2022.

Deprez, G. 2019. The destruction of nude images, last access: 31 may 2022.

Deprez, G. 2020. Cover up that bosom which i can't endure to look on, last access: 31 may 2022.

Eck, B. A. 2001. Nudity and framing: Classifying art, pornography, information, and ambiguity. *Sociological Forum* 16(4):603–632.

Elkin-Koren, N. 2020. Contesting algorithms: Restoring the public interest in content filtering by artificial intelligence. *Big Data & Society* 7(2):2053951720932296.

Faust, G. 2017. Hair, blood and the nipple. In *Digital Environments*. transcript Verlag. 159–170.

Guckelsberger, C.; Kantosalo, A.; Negrete-Yankelevich, S.; and Takala, T. 2021. Embodiment and computational creativity. *arXiv preprint arXiv:2107.00949*.

Hunt, E. 2021. Vienna museums open adult-only onlyfans account to display nudes, last access: 31 may 2022.

Ibrahim, Y. 2017. Facebook and the napalm girl: Reframing the iconic as pornographic. *Social Media + Society* 3(4):205630511774314.

James, P. 2014. 8 reasons why social media is decimating art and literature, last access: 31 may 2022.

Kosko, B. 1999. *The fuzzy future: from society and science to heaven in a chip*. Harmony.

Kulesz, O. 2018. Culture, platforms and machines: the impact of artificial intelligence on the diversity of cultural expressions. *Intergovernmental committee for the protection and promotion of the diversity of cultural expressions*.

Manovich, L. 2018. *AI aesthetics*. Strelka Press Moscow.

Montaner, J. 1999. Arquitectura y crítica. *Gustavo Gili*.

Mulvey, L. 1975. Visual pleasure and narrative cinema. *Screen* 16(3):6–18.

Oliver, K. 2017. The male gaze is more relevant, and more dangerous, than ever. *New Review of Film and Television Studies* 15(4):451–455.

Olszanowski, M. 2014. Feminist self-imaging and instagram: Tactics of circumventing sensorship. *Visual Communication Quarterly* 21(2):83–95.

Patridge, S. 2013. Exclusivism and evaluation: Art, erotica and pornography. In *Pornographic Art and the Aesthetics of Pornography*. Palgrave Macmillan UK. 43–57.

Polaine, A.; Street, S.; and Paddington, S. 2005. Lowbrow, high art: Why big fine art doesn't understand interactivity.

Pollen, A. 2021. Pubic hair, nudism and the censor: The story of the photographic battle to depict the naked body.

Ramirez, J. 1998. *Art History and critique: faults (and failures)*. F. Cesar Manrique.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17:76–99.

Saunders, R., and Bown, O. 2015. Computational social creativity. *Artificial life* 21(3):366–378.

Schroeder, J. 2021. Reinscribing gender: social media, algorithms, bias. *Journal of Marketing Management* 37(3-4):376–378.

Sezen, D. 2020. Machine gaze on women: How everyday machine-vision-technologies see women in films. In *Female Agencies and Subjectivities in Film and Television*. Springer International Publishing. 271–293.

Sharlow, S. 2015. Death of an artist: How social media is ruining creativity, last access: 31 may 2022.

Vasari, G. 1550. *Le vite de' più eccellenti architetti, pittori, et scultori italiani, da Cimabue insino a' tempi nostri*.

Vasilaki, M. 2010. Why some pornography may be art. *Philosophy and Literature* 34(1):228–233.

Vattimo, G. 1988. *The End of Modernity: Nihilism and Hermeneutics in Post-Modern Culture*. Polity Press in Association with B. Blackwell.

Ventura, D. 2017. How to build a cc system. In *ICCC*, 253–260.

Volkers, R. 2020. Perverse media: How instagram limits the potential of feminist art, last access: 31 may 2022.

Walker, M., and Matsa, K. E. 2021. News consumption across social media in 2021, last access: 19 may 2022.

Wang, X.; Cheng, F.; Wang, S.; Sun, H.; Liu, G.; and Zhou, C. 2018. Adult image classification by a local-context aware network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2989–2993.