Information Research - Vol. 30 No. iConf (2025)

Safety anchors and deflected desires: generative AI and the production of sexualities

Daniel Carter
DOI: (https://doi.org/10.47989/ir30iConf47569

Abstract

Introduction. Many images generated by AI appear sexual in nature, and a great deal of attention is paid in computer science to preventing systems from producing some forms of sexual content. Yet, despite the fact that sex seems to be everywhere in generative AI, we are saying little about how AI might intervene in our sexualities.

Approach. This exploratory essay argues for the need to consider sexuality in relation to generative AI and draws on broad arguments contained in Michel Foucault's work on sexuality to set the stage for future work. Drawing on Rieder's concept of algorithmic techniques, it considers current methods of prohibiting the generation of sexual content and uses these as a base for discussing the ways that AI might play a role in the formation of sexualities.

Conclusion. This essay aims to provoke future work by sketching out some aspects of generative AI that might be relevant to studies of sexuality. Primarily, it highlights the processes of evasion and denial as potentially drawing attention to and concretizing material or concepts that are considered taboo. These observations are not intended to be validated or generalizable; they are instead intended only to illustrate the potential of future work.

Introduction

This essay is prompted by the curious situation in which much of the actual output of generative AI systems represents idealized, sexualized female bodies while at the same time researchers in the field of machine learning produce volumes of work intending to prevent these same systems from producing images that might be seen as overly sexual, inappropriate or pornographic. We are, in a sense, saying a great deal about sex—whether through the production of images or through the production of texts that intend to circumvent such production—while at the same time saying quite little about the ways in which generative AI intervenes in our sexualities. While considerable research exists on pornography from the perspective of media reception, the technical specifics of generative AI warrant asking these questions about sexuality from a perspective that accounts for what Rieder (2020) calls '*algorithmic techniques*.' This short essay sets the ground for future work in this area by placing generative AI in conversation with Foucault's work on sexuality, for example by prompting questions related to the experiences of evasion and denial.

To orient this exploratory essay, I first review several broad points from Michel Foucault's work on sexuality before summarizing the engagement that media studies and information studies has had with the topic. I then highlight thematic algorithmic techniques related to the prohibition of undesired content in AI systems, discussing how these might provoke future work.

Foucault's theory of sexuality

In searching for a theoretical grounding on sexuality as a basis for examining generative AI systems, I turn to Michel Foucault's work due to its profound influence and its analysis of the complex networks of institutions, discourses, and technologies that constitute and regulate subjectivities. Particularly relevant is Foucault's attention to the dynamics of regulation, control, and prohibition concerning sex and sexual content, primarily developed in '*the history of sexuality*' (2012a, 2012b). Rather than attempting a comprehensive survey of theoretical approaches to sexuality or an exhaustive account of Foucault's contributions and their ongoing reception, this brief review adopts an abductive approach suitable for an exploratory essay, intending to highlight directions that may prove fruitful for future work. I focus on two key arguments from Foucault's work that are especially generative for understanding and conceptualizing algorithmic techniques related to the formation of sexualities:

First, a person's sexuality is not an innate characteristic that is either repressed or liberated. Instead, it is a set of practices, desires, and identities produced through interaction with a historically contingent apparatus composed of discourses, institutions, laws, and social norms that define some identities and behaviors as acceptable while marking others as inappropriate or taboo. By critiquing the repressive hypothesis—which would posit, for example, a reduction in sex and talk of sex during repressive periods such as the Victorian era—Foucault argues that prohibiting certain sexual acts or identities does not erase them from society but instead gives rise to a proliferation of discourses about them, which reinforce their presence and create new possibilities for individuals to identify as, for example, deviant or sick.

Second, Foucault claims that, since the 18th century, the formation of individual sexualities has been closely tied to '*sex-writing*'—the ways in which sexuality is inscribed, whether through professional texts, personal confessions, or broader mechanisms of discourse and governance. This 'endless discourse' on sex has become a central means by which individuals come to understand and express their own desires. Alongside professional, scientific, and juridical discourses, the rise of personal confessions, diaries, and therapeutic dialogues has compelled individuals to '*speak of their sex*,' turning private thoughts and behaviors into objects of scrutiny and regulation. By narrating their intimate experiences, individuals respond to and internalize societal norms, allowing power to operate not just from above but within the very processes of

self-description, self-disclosure and self-formation. Consequently, sexuality becomes both a target and an instrument of power, as the act of speaking about sex—whether in private or public—along with the various ways in which it is regulated and categorized, contributes to the continuous production of sexual identities, whether aligned with or resistant to dominant norms.

Sexuality and sexual content in information systems and media studies research

This section presents a brief review of the ways that sexuality and sexual content have been approached in the fields of information systems and media studies. In the context of generative AI systems, both fields of knowledge are relevant, as media studies, through its engagement with pornography, has had a greater historical interest in people's interactions with sexual content, while information systems researchers have an obvious engagement with technical features of systems such as algorithms and taxonomies, despite only less historical interest in how these influence the formation of sexualities.

Regarding sexuality, early approaches in both media studies and information systems often treat the concept as an innate and stable characteristic. In media studies, this perspective is evident in research focused on the biological or psychological effects of sexual content, where sexuality is framed as a natural drive influenced by external stimuli. For example, studies on the impact of pornography often assume a direct, measurable effect on behavior, operating on the premise that sexual desires are inherent and can be either stimulated or suppressed by media exposure. Similarly, in the field of information systems, early research on online behavior frequently relied on user profiling that treated sexuality as a fixed variable.

Following the turn, in media studies, to consider more individual and situated uses of media (e.g., Ang, 1991; Jenkins, 2005; Radway, 1984), more recent research has explored how sexuality is negotiated through or produced in conjunction with available media. For example, Spigel's (1992) work on postwar television and domesticity demonstrates how TV programs played a role in shaping gendered and sexual norms within the private sphere, reinforcing traditional ideas of family and heteronormativity. Similarly, the field of information systems has examined how technical features such as platform affordances or taxonomic categories shape the representation of sexualities (Light, 2007; Light et al., 2008; Ruberg & Ruelos, 2020; Watson, 2020) and, to a lesser extent, the enactment of desire (Paasonen et al., 2019). Still, there is far more work in both fields that focuses on gender than on sexuality, perhaps mirroring the pattern discussed, below, of researchers avoiding the topic of pleasure, especially as it relates to sexual content.

Research on sexual content, the most prominent example of which being pornography, follows a similar course, with early interest (e.g., Dworkin, 1981; MacKinnon, 1996) focusing on the effects of sexual content. This work, much of which had the goal of enacting legal restrictions, argued that pornography was inherently chauvinistic and that it had the effect of encouraging oppressive and violent versions of male sexuality. Later theorists such as Williams (1999) critiqued these attacks on pornography as overly simplistic, arguing that they assumed direct and consistent effects rather than acknowledging individuals' various uses and negotiations of sexual content.

While there has been less attention given to pornography in the information sciences, scholars have shown a keen interest in the moderation of sexual content from online platforms. Research in this area has critically examined how platforms enforce content guidelines from the perspective of both labor (Roberts, 2019) and algorithms (Gillespie, 2022), with the bulk of attention paid to the processes that lead sexual content to disappear from digital spaces.

While studies of algorithms and content moderation do examine the ways that power acts to shape the content that is made available, much less work has enquired into the ways that the presence or non-presence of sexual content shapes sexualities, specifically in relation to the experience of

pleasures and identification of objects of pleasure, although notable exceptions to this trend exist (e.g., Keilty, 2012; Keilty & Leazer, 2018). Indeed, researchers in the field of human-computer interaction (HCI) have repeatedly remarked on the paradox of pornography's ubiquity and the lack of research that asks about the gratifying uses of such content (Blythe & Jones, 2004; Su et al., 2019). It's also notable that, while HCI researchers have been more responsive to calls to focus on pleasure, they have been far more likely to consider objects such as sex toys (Bardzell & Bardzell, 2011; Hua et al., 2022) or sex robots (Fosch-Villaronga & Poulsen, 2021; Su et al., 2019) than to study sexual images or videos.

As a consequence, the discourse around sexual content and sexuality remains largely constrained to concerns over harm and regulation, rather than expanding to consider the formation of diverse experiences of desire. While information systems research has only started to engage in this area, its contributions are especially relevant given the technical differences between older media technologies and generative AI systems that make accessible the production of sexual content in addition to its widespread dissemination. While it seems likely that the most common interaction that people have is viewing AI-generated content (as opposed to generating such content), the possibility of future technologies that use generative AI to create personalized experiences warrants an investigation, now, into the ways that sexualities form in interaction with what Rieder (2020) refers to as algorithmic techniques. As Rieder argues, these techniques exist between software studies' interest in specific code manifestations and media studies' focus on content and products. As standardized implementations of knowledge that exists in archives such as scholarly papers and online discussion forums, algorithmic techniques offer a way in to the black box of software and present researchers with a tool with which to comment on the functioning of commercial software without access to its code.

In the following sections, I draw on basic tenets of Foucault's theory of sexuality and ask how the algorithmic techniques underlying current generative AI systems might influence the formation of sexualities and the identification of objects of desire. Appropriate for a short, exploratory essay, these remarks are intended to function only as short provocations for future work, tracing paths that could be followed rather than arguing for validity or generalisability.

Erased concepts, safety anchors, coded confessions, and deflected desires

I focus in this section on algorithmic techniques that attempt to preclude the creation by image-generating AI systems of content that is deemed unacceptable and that operate on pre-trained models; that is, these techniques intervene in models that have already been trained and are capable of producing content deemed sexual or undesirable, as opposed to training a model that has been explicitly designed to exclude such information from its training data. While there are various broad categories of relevant techniques, including post-generation techniques that classify or censor content after it is produced, I focus here on technique that identify specific vectors or regions within the model's latent space—mathematical structures that correspond to the representation of certain types of content, such as sexual or undesirable material and which can then be manipulated to guide the generation process away from producing such content. While specific implementations of this technique vary in their technical details in ways that certainly affect the produced content, the broad concepts of '*safe*' and '*forbidden*' zones, introduced below, offer provocative ways to consider the formation of sexualities.

For example, Schramowski et al.'s (2023) safe latent diffusion (SLD) approach identifies a '*general inappropriate safety concept*' using the text prompt, '*hate*, *harassment*, *violence*, *suffering*, *humiliation*, *harm*, *suicide*, *sexual*, *nudity*, *bodily fluids*, *blood*, *obscene gestures*, *illegal activity*, *drug use*, *theft*, *vandalism*, *weapons*, *child abuse*, *brutality*, *cruelty*,' and notes that this approach is inspired by existing techniques used in commercial models such as DALL-E. At each step of the

diffusion process, SLD modifies the vector represented by the original prompt in order to move it away from the vector representing the '*general inappropriate safety concept*' and toward a safety anchor or point in latent space that is dynamically calculated in relation to the entered prompt. An exemplary result of the process, given in Schramowski et al.'s paper, is the generation of an image of a woman wearing a dress when, without intervention, the model would have generated an image of a nude woman.

A related approach, concept erasure, attempts to identify broad concepts within a pre-trained model in order to remove them from the model weights. For example, Gandikota et al. (2023) demonstrate that concepts such as nudity, artistic styles such as that of Van Gogh or entire classes of objects such as cars can be erased from a pre-trained model, resulting in a modified model that does not include the information to produce relevant images. This approach differs from the application of safety anchors, described above, in that the authors argue that, rather than a model that could potentially produce unwanted images but does not, they produce a model that does not have the ability to produce such images.

Relevant adversarial techniques in this area include the modification of prompts to avoid prohibited terms while still producing a vector that results in the generation of an image that would be considered inappropriate. Tsai et al. (2024), for example, demonstrate a technique for generating text prompts that correspond to concepts that have been suppressed (by a technique described above, for example). The assumption made is that models contain implicit text-concept associations that can be discovered, and the authors demonstrate that prompts including strings such as '*nucropsexynipples*', '*nakedcougarscenery*', as well as terms whose semantic meaning is more inscrutable, can be used to evade the safety mechanisms in current software.

Notably, Tsai et al. argue that their technique is practical and briefly describe a use case in which their tool would be run offline in order to translate prompts that would then be entered into an existing software product to evade censoring. This note brings attention to the experience of users who attempt to, and perhaps succeed in, generating sexual content that is marked as undesirable by current systems. Here I briefly focus on two experiences that appear provocative in relation to Foucault's work on sexuality.

First, the translation of one's desires into a new and obscured language, such as with Hsu et al.'s technique, might be seen as a kind of technologically mediated '*sex writing*', drawing attention to one's existing desires and their taboo status. This process exists in a similar form in observed practice, as users have noted that phonemes and similar word parts appear to have consistent meaning and can be used to generate desired images (e.g., 'Odd Little Spots in Midjourney Latent Space,' 2023). As a form of sex writing, or of externalizing one's desires, this translation inserts both new rituals and also new semantics into the formation of sexualities, producing phrases that in many ways behave as magic spells: sequences of largely nonsensical words that produce an effect in the world that is otherwise prohibited. While power operates at the level of law, science and industry to forbid the depiction of nude bodies, possessing the knowledge and capacity to circumvent that power gives to the produced images a special status as contraband. Following a media studies approach such as would be advocated by McLuhan (2013) and which might argue for the transference of erotic desire from the content of a pornographic film to its media qualities (e.g., cuts and transitions), in the ritual of translating an expressed desire into an esoteric language that has the power of evading censorship and summoning forbidden images, we might also imagine a similar transference.

Second, Foucault's insistence that repression does not banish desire but instead constructs it suggests the merits of attending carefully to the moment at which a user of an AI system is denied what was requested. Here the dynamics of the algorithms described above seem quite relevant, as the experience is specifically of expecting one thing and — rather than receiving nothing —

receiving the thing that was requested, with as much fidelity to the original request as possible but with the exception of a concept such as nudity removed. Relevant pre-AI experiences that serve as useful cognates here include, perhaps, searching for a pornographic image and seeing blurred rectangles returned or watching a film in anticipation of a sexual scene that does not appear. As Foucault would argue, such experiences do not erase the object of desire; rather, they likely concretize the object and the user's conception of self, creating a moment of drama in which one waits for their request to be revealed as acceptable or taboo.

While these brief sketches warrant greater elaboration, and engagement with both theories of sexuality as well as empirical data related to users' interactions with AI systems, they suggest the directions that such work might take and the novel ways that such work might expand the study of information systems to include how these impact the formation of sexualities and the selection of objects of desire.

Conclusion

This essay aims to provoke future work by sketching out some aspects of generative AI that might be relevant to studies of sexualities. Primarily, it highlights the processes of evasion and denial as drawing attention to material or concepts that are considered taboo. These observations not intended to be validated or generalizable; they are instead intended only to illustrate to potential of future work.

About the author

Daniel Carter is an Associate Professor in the School of Journalism and Mass Communication at Texas State University. They can be contacted at dcarter@txstate.edu.

References

Ang, I. (1991). *Desperately Seeking the Audience*. Routledge.

Bardzell, J., & Bardzell, S. (2011). Pleasure is your birthright: Digitally enabled designer sex toys as a case of third-wave HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 257–266.

Blythe, M., & Jones, M. (2004). Human computer (sexual) interactions. *Interactions*, *11*(5), 75–76.

Dworkin, A. (1981). *Pornography: Men Possessing Women* (First Perigee Printing edition). The Women's Press Ltd.

Fosch-Villaronga, E., & Poulsen, A. (2021). Sex robots in care: Setting the stage for a discussion on the potential use of sexual robot technologies for persons with disabilities. *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 1–9.

Foucault, M. (2012a). *The History of Sexuality: An Introduction*. Knopf Doubleday Publishing Group.

Foucault, M. (2012b). *The History of Sexuality, Vol. 2: The Use of Pleasure*. Knopf Doubleday Publishing Group.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D. (2023). Erasing concepts from diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media + Society*, 8(3), 20563051221117552. https://doi.org/10.1177/20563051221117552

Hua, D. M., Jones, R., Bardzell, J., & Bardzell, S. (2022). The Hidden Language of Vibrators: A Politico-Ontological Reading. *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 400–414.

Jenkins, H. (2005). *Textual Poachers: Television Fans & Participatory Culture*. Routledge : Taylor & Francis. http://www.UTXA.eblib.com/EBLWeb/patron?target=patron&extendedid=P 180969 0&

Keilty, P. (2012). Embodiment and desire in browsing online pornography. *Proceedings of the 2012 iConference*, 41–47. https://doi.org/10.1145/2132176.2132182

Keilty, P., & Leazer, G. (2018). Feeling documents: Toward a phenomenology of information seeking. *Journal of Documentation*, 74(3), 462–489. https://doi.org/10.1108/JD-09-2016-0113

Light, B. (2007). Introducing masculinity studies to information systems research: The case of Gaydar. *European Journal of Information Systems*, 16(5), 658–665.

Light, B., Fletcher, G., & Adam, A. (2008). Gay men, Gaydar and the commodification of difference. *Information Technology & People*, 21(3), 300–314.

MacKinnon, C. A. (1996). *Only Words*. Harvard University Press.

McLuhan, M. (2013). *Understanding Media: The Extensions of Man*. Gingko Press.

Odd little spots in Midjourney latent space. (2023, March 27). *Ceoln*. https://ceoln.wordpress.com/2023/03/27/odd-little-spots-in-midjourney-latent-space/

Paasonen, S., Light, B., & Jarrett, K. (2019). The dick pic: Harassment, curation, and desire. *Social Media+ Society*, 5(2), 2056305119826126.

Radway, J. (1984). *Reading the romance: Women, patriarchy, and popular literature*. University of North Carolina Press.

Rieder, B. (2020). *Engines of order: A mechanology of algorithmic techniques*. Amsterdam University Press.

Roberts, S. T. (2019). *Behind the Screen*. Yale University Press.

Ruberg, B., & Ruelos, S. (2020). Data for queer lives: How LGBTQ gender and sexuality identities challenge norms of demographics. *Big Data & Society*, 7(1), 2053951720933286. https://doi.org/10.1177/2053951720933286

Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2023). Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Spigel, L. (1992). *Make Room for TV: Television and the Family Ideal in Postwar America*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/M/bo3624766.html

Su, N. M., Lazar, A., Bardzell, J., & Bardzell, S. (2019). Of dolls and men: Anticipating sexual intimacy with robots. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(3), 1–35.

Tsai, Y.-L., Hsu, C.-Y., Xie, C., Lin, C.-H., Chen, J.-Y., Li, B., Chen, P.-Y., Yu, C.-M., & Huang, C.-Y. (2024). *Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models?* (No. arXiv:2310.10012). arXiv. https://doi.org/10.48550/arXiv.2310.10012

Watson, B. M. (2020). 'There was Sex but no Sexuality*:' Critical Cataloguing and the Classification of Asexuality in LCSH. *Cataloging & Classification Quarterly*, 58(6), 547–565. https://doi.org/10.1080/01639374.2020.1796876

Williams, L. (1999). *Hard Core: Power, Pleasure, and the "Frenzy of the Visible"*, *Expanded edition* (First Edition). University of California Press.