Censorship, Artificial Intelligence, and AI Literacy: Responding to New Challenges in an Enduring Agenda for Libraries and Librarianship

Michael Ridley, Librarian Emeritus, University of Guelph
mridley@uoguelph.ca

avram anderson, Collection Management Librarian,
California State University, Northridge
avram.anderson@csun.edu

April 2025

Currently under consideration for publication in the Journal of Radical Librarianship

Abstract

This paper focuses on AI censorship, an under addressed aspect of AI risk that intersects with the foundational library tenets of information literacy and intellectual freedom. AI censorship is a form of "automated censorship in which AI systems are used to selectively suppress or block specific types of information, content, or voices deemed undesirable to those controlling the AI." This paper examines AI censorship in the context of existing threats and library principles. It explores specific techniques and methods of AI censorship. Lastly, it recommends adopting a critical AI literacy perspective that includes a political dimension essential to understanding AI censorship.

Keywords: censorship, artificial intelligence, critical AI literacy

Introduction

"Censorship has been, is, and will continue to be one of the single most important issues for librarians."[1] Librarians and library associations have historically taken up the challenge of censorship and remain vigilant as threats, such as book bans, are increasing.[2] Anti-censorship initiatives rest in the commitments of librarianship to intellectual freedom, inclusive spaces, and service to diverse populations. Those commitments face new challenges in the digital world. Specifically, the rapid rise and impact of generative AI and its capability to censor "with increasing ease, and at great scale and speed" has been largely overlooked.[3] The "repressive power" of AI is a new front for censorship and one that requires libraries to broaden their understanding of AI and to develop policies and programming to address the threats.[4]

The risks associated with AI, and large language models (LLMs) specifically, have been widely documented.[5] Misinformation, disinformation, deepfakes, bias, discrimination,

cybersecurity, and a host of other threats and deficiencies require concerted attention.[6] The library and information science (LIS) community has been responsive in addressing the implications of AI for libraries and librarianship.[7] In response, the rapid and widespread design and delivery of AI literacy as a core component of information literacy curriculum is notable.[8]

This paper will focus on AI censorship, an under addressed aspect of AI risk that intersects with the foundational library tenets of information literacy and intellectual freedom. AI censorship is a form of "automated censorship in which AI systems are used to selectively suppress or block specific types of information, content, or voices deemed undesirable to those controlling the AI."[9]

The American Library Association defines censorship as a "change in the access status of material, based on the content of the work and made by a governing authority or its representatives. Such changes include exclusion, restriction, removal, or age/grade level changes."[10] The Canadian Federation of Library Associations links censorship and intellectual freedom calling for libraries to "provide, defend and promote equitable access to the widest possible variety of expressive content and resist calls for censorship and the adoption of systems that deny or restrict access to resources."[11] The use of AI to enable censorship to deliberately exclude, restrict, suppress, and remove information "has the potential to reshape the information landscape by omission … with effects less visible than the proliferation of disinformation or online abuse."[12]

This paper will position AI censorship in the context of existing threats and library principles. It will explore specific techniques and methods of AI censorship. Lastly, it will recommend adopting a critical AI literacy perspective that includes a political dimension essential to understanding AI censorship.

Context

The call to action is frank: "a dangerous authoritarian vision of the future of AI is taking shape. The time has come for democratic actors to mount a serious response."[13] While AI censorship is typically associated with foreign, authoritarian governments, H. Akın Ünver notes that the

> technology monopolies of Google, Baidu, Alibaba, Amazon, YouTube, Tencent, Facebook and others, coupled with Silicon Valley-style PR brinkmanship culture will likely lead to more dangerous and unnerving developments in algorithmic politics in comparison to what democratic or authoritarian states may or may not do with A.I.[14]

The Freedom House report, *The Repressive Power of Artificial Intelligence*, discusses the continued increase in the use of these technologies to suppress freedom of expression. Global internet freedom, which Freedom House measures through obstacles to access, limits on content, and violations of user rights, has continued to decline. This is not limited to authoritarian governments: "States that have long been

defenders of internet freedom have imposed censorship or flirted with proposals to do so, an unhelpful response to genuine threats of foreign interference, disinformation, and harassment."[15]

At stake is what Dan Hendrycks and Mantas Mazika describe as "value lock-in" where

> strong AI imbued with particular values may determine the values that are propagated into the future. Some argue that the exponentially increasing compute and data barriers to entry make AI a centralizing force. As time progresses, the most powerful AI systems may be designed by and available to fewer and fewer stakeholders. This may enable, for instance, regimes to enforce narrow values through pervasive surveillance and oppressive censorship.[16]

Yuval Noah Harari, in his discussion of democracy, authoritarianism and AI, is blunt about the effects of AI centralization: "a single archive makes censorship easy."[17]

User education, specifically AI literacy, is often identified to mitigate AI risks. However, despite the concern about AI censorship, recent reviews of AI literacy failed to identify censorship as an element.[18]

The Role for Libraries and Librarians

"Censorship is an evil thing. In accepting it, in compromising, in 'playing it safe,' the librarian is false to the highest obligations of his [sic] profession."[19]

Leon Carnovsky (1902-1975)

Gary Marchionini explicitly identifies the role of librarians in contesting issues like AI censorship because of librarians' experience

> in advocating for and providing training for literacy to all, our incorporation of new media in public collections, our opposition to censorship, our expertise in critical assessment of knowledge assets, and our devotion to equal access have prepared us to take leadership roles as new products and outcomes of GenAI applications affect individuals and society.[20]

Martin Frické is more emphatic regarding this role suggesting that librarians act as "sentries ... protecting patrons … [and] wrestling with censorship."[21]

With their continuing focus on equity and access, librarians should be at the forefront of AI adoption in our everyday lives, championing AI literacy and critical thinking that are essential for understanding AI's limitations, biases, and ethical implications.

Techniques and Examples of AI Censorship

AI censorship differs from what we might call "classic" censorship techniques. Book banning, for example, is transactional, visible, and contestable. There is a clear process undertaken by perpetrators and witnessed (and potentially resisted) by others. Book

censorship happens in clear sight. AI censorship, as with most algorithmic processes, is largely invisible, conducted by code not people, and difficult to resist. AI censorship happens in the digital shadows.

An important characteristic of AI censorship follows from the generative AI models it often uses. Generative AI, particularly in its implementation in chat systems, is not a conventional online search system but rather an encompassing environment that creates a specific and persuasive worldview.[22] The chat interface and modality encourage, even rely on, a personal relationship between the user and the system. It feels like another person; it feels natural. The rise of AI companions, such as Replika, epitomize this effect.[23] The result is an over trust in AI, a form of "automation bias."[24] We are predisposed to believe the output of the system.

a) Authoritarian AI Systems

The most blatant use of AI censorship is in AI systems built and deployed by authoritarian states for use primarily within their own countries. These systems, typically chatbots, are trained on highly selective datasets and have various forms of guardrails to further suppress and conceal unapproved topics or issues. As Sarah Zheng notes, "with artificially intelligent chatbots, censorship comes built-in."[25]

Systems designed and deployed in Russia and China are examples of this. In these cases, requests for certain topics are ignored or deflected, keywords are suppressed, and attempts to jailbreak are monitored. For example, in China, Baidu's Ernie bot refused to respond to inquiries about Tiananmen Square, the Russian Kandinsky 2.1 neural network not only obfuscated any request regarding Ukraine it also blocked the use of "gays", "lesbians", and "LGBT," and Alice, from the Russian Yandex company, provided only a vague response to questions about Alexey Navalny.[26]

Ernie-ViLG (a text-to-image AI from Baidu) complies with the Chinese list of "sensitive words" censoring requests for terms such as "revolution" and "Tiananmen Square."[27] A similar Russian image creation system simply responds with flowers when prompted with the word "protests."[28] However, these restrictions can be uneven: the Ernie bot blocked requests in Chinese but not in English.[29]

The R1 models released by China's DeepSeek in late 2024 have become some of the most powerful and widely used generative AI systems. Early adopters quickly discovered the model's compliance with China's restrictive generative AI rules including pro-Chinese Communist Party (CCP) positions, historical distortions, deflections, and refusals to respond.[30] With DeepSeek, AI censorship has gone mainstream.[31]

b) Content Moderation as AI Censorship

While the AI censorship actions of authoritarian governments are obvious, it is more complex and controversial to view the widely deployed techniques of content moderation as potential forms of AI censorship. These government and corporate practices have been called “censorship by proxy.”[32]

Content moderation is defined as “the automated monitoring and filtering of user-generated content by online platforms, guided by their legal obligations, in-house content policies and commercial objectives.”[33] Content moderation is employed by most AI chat systems (e.g., ChatGPT, Gemini, CoPilot, and others). It enables guardrails to prevent illegal, malicious or offensive output. Based on notions of community standards, it prevents expressions of hate or violence, pornography, and child exploitation. For the most part, users accept and support such restrictions, even if they are unaware of how it works or to the extent it influences the system responses.

While there are widely adopted principles of content moderation (e.g., The Santa Clara Principles on Transparency and Accountability in Content Moderation, 2021), the specific definition of “community standards” is often vague and largely a product of each platform’s judgement not the result of government regulation or societal deliberations.[34] Describing content moderation as “algorithmic censorship,” Jennifer Cobbe notes that

> the emergence of algorithmic censorship as a primarily commercially driven mode of control undertaken by social platforms is an undesirable development that empowers platforms by permitting them to more effectively align both public and private online communications with commercial priorities while in doing so undermining the ability of those platforms to function as spaces for discourse, communication, and interpersonal relation.[35]

At issue is the line where content moderation become censorship. Efforts to suppress genuine concerns might inadvertently suppress otherwise legitimate interests and ideas.[36] Experience with social media illustrates that removal of data through content moderation disproportionally effected political conservatives, transgender people, and Black people. However,

> conservative participants’ removals often involved harmful content removed according to site guidelines to create safe spaces with accurate information, while transgender and Black participants’ removals often involved content related to expressing their marginalized identities that was removed despite following site policies or fell into content moderation gray areas.[37]

As Summer Lopez documents, popular chatbots and text-to-image systems have broad restrictions and prohibitions:

> OpenAI’s usage policies include a litany of forbidden uses of ChatGPT, including “fraudulent or deceptive activity,” alongside everything from gambling and weapons development to adult content creation. In addition to the company’s existing terms of service, Google’s generative AI prohibited use policy groups all barred activities under “dangerous, illegal, or malicious activity,” “content

> intended to misinform, misrepresent, or mislead," and "sexually explicit content." Microsoft supplements the code of conduct in its services agreement with additional provisions for its Bing chat and image creator, which state the user must not generate content that is illegal, harmful, or fraudulent.[38]

When Google's Gemini was asked (on August 28, 2024), about whether it trains on different data for different jurisdictions it noted not just the use of local datasets but the use of "filtering and cleaning" to "remove irrelevant or harmful content" and "adapting training algorithms" or fine-tuning to respond to local requirements. While ChatGPT says it doesn't train on different data for different jurisdictions, it did acknowledge (on August 28, 2024) that "ChatGPT does not provide specific LGBTQ+ resources for Middle Eastern countries due to the potential legal and cultural sensitivities surrounding LGBTQ+ issues in many of these regions." In other words, in both cases, these systems deliberately censor materials.

The Santa Clara principles regarding content moderation assert that "state actors must not exploit or manipulate companies' content moderation systems to censor dissenters, political opponents, social movements, or any person."[39] However, a report from Freedom House found that 45 countries had forced platforms to remove content from their sites or systems with many countries "obliging platforms to use machine learning to comply with censorship rules, governments are effectively forcing them to detect and remove banned speech more efficiently."[40]

c) AI Censorship Allegations and Responses

AI censorship has been alleged or demonstrated in some of the most widely used AI-based services in North America. Google and Meta have been accused of censoring right-leaning views for domestic consumption.[41] The algorithms that drive TikTok "consistently amplify pro-CCP content and suppress anti-CCP narratives."[42]

Mark Zuckerberg, CEO of Meta, in a letter to the Committee on the Judiciary in the US Congress, outlined what he viewed as censorship requests and pressures from the US Executive Branch. That same Committee alleges that the National Science Foundation has developed "artificial intelligence (AI)-powered censorship and propaganda tools that can be used by governments and Big Tech to shape public opinion by restricting certain viewpoints or promoting others."[43] Recently Zukerberg removed third-party content moderation from Facebook and Instagram arguing it infringes of free speech.[44] Notably this change only pertains to US users and appears not to impact Meta's AI models.

However, in response to concerns about content moderation enabling censorship, there has been a proliferation of "uncensored" chatbots, such as Grok with no guardrails.[45] FreedomGPT was created, according to its creators, "to answer any question without censorship, judgment, or 'post-inference bias.'" The result has been called "an equal opportunity offender" in its unmoderated responses. John Arrow, the creator of FreedomGPT defends the system, suggesting by analogy that there is "no expectation for the pen to censor the writer."[46]

The availability of user-friendly development tools has facilitated the proliferation of custom chatbots, such as FreedomGPT and those derived from OpenAI's ChatGPT. These chatbots are valuable for small organizations wanting to fine-tune on local datasets. For many, this is an important way to create AI that responds to specific business or organizational objectives. For others, however, it is a way to create systems that deliberately suppress information for their users.

The concern is where these custom GPTs become a mandated tool for groups of users (e.g., employees, students, religious adherents) or a preferred tool (e.g., political parties, extremist groups). Designed as suggested, these systems would be created not simply for organizational objectives but additionally to intentionally suppress, rather than contest, opposing views. Clearly many of these local GPTs are not involved in censorship. It is where GPTs are designed to exclude or suppress, and where users lack a more holistic alternative that censorship emerges.

d) Training Data and AI Censorship

Training data is central to the scope and effectiveness of generative AI, and we have shown how highly selective datasets can train AI to function as a censorship tool. Training data "has direct implications for model behavior and who is likely to be empowered and disempowered by applications built on top of those models."[47] Even datasets widely believed to be appropriate can have embedded issues. Many LLMs are pre-trained using the extensive Common Crawl dataset.[48] The dataset includes over 320 websites that obey Chinese censorship laws. As a result, these LLMs models have not only learned censored content, but they have perpetuated this content in downstream uses of these models.[49]

One demonstration of the effect of this training is an experiment which assessed the differences between a set of prompts to ChatGPT3.5 Turbo in Simplified Chinese characters (used primarily in mainland China where censorship is rigorous) and Traditional Chinese characters (used primarily in the more liberal Taiwan and Hong Kong).

> The Simplified responses tended to gloss over or even ignore details concerning the CCP and Xi Jinping's human rights records, territorial disputes, and other prohibited topics in China, whereas the Traditional responses were more critical of the regime, outlined human rights violations in much more depth, and were more sympathetic to victims of oppression.[50]

ChatGPT was, in part, pre-trained on Common Crawl. Prompting in a different character set was all that was needed to elicit responses governed by Chinese censorship rules. Similarly, ChatGPT gives more positive responses about Mao Zedong when queried in Chinese than in English.[51]

Common Crawl faces accusations of bias and censorship resulting from the methodology used to determine which domains to include in the crawling and indexing processes.[52] The current approach relies on the calculation of "harmonic centrality," a

significance score which considers the number of direct and indirect links that exist from other domains. Relying on link-based ranking has been shown to be problematic because content from niche or marginalized identities will likely have a lower domain centrality score due to having fewer inbound links and less mainstream coverage. This results in lower visibility and reduced inclusion in the dataset.[53]

However, it should be noted that the originating goal of Common Crawl was not to train LLMs but rather to act as a source for research and development:

> Common Crawl wants its data to contain problematic content to enable open-ended research and innovation, but it does not want to take responsibility for annotating it. For LLM development, this is a problematic starting point that requires careful consideration for how Common Crawl's data is filtered before any model training.[54]

The most frequently used versions of filtered datasets for training often try to remove pornographic content. However, the way that that data is classified or categorized can remove legitimate content related to sexual minority groups if a more nuanced approach isn't developed to take this into consideration. Research on determining demographic identities that have been excluded from training datasets due to filtering has shown that "mentions of sexual orientations (lesbian, gay, heterosexual, homosexual, bisexual) have the highest likelihood of being filtered out" and "documents associated with Black and Hispanic authors and documents mentioning sexual orientation are significantly more likely to be excluded."[55]

Despite the importance of training data to AI, "comparatively little attention has been paid to how and why machine learning datasets have been created, what and whose values influence the choices of data to collect, the contextual and contingent conditions of their creation."[56] Orr & Crawford call this "the messy and contingent realities of dataset preparation" which opens the door to datasets as a tool for AI censorship.[57]

e) Technical Building Blocks and Adversarial Interventions

So far, we have discussed censorship in the context of an entire systems, whether commercial, governmental, or organizational. There are narrower, more technical approaches that involve building blocks or interventions.

Model diversion, data poisoning, adversarial attacks, and jailbreaking are methods to alter the output of the AI model to disrupt and to effectively censor information. The techniques work differently and at different stages of the model's training and deployment, but the result is to suppress certain ideas and amplify (or distort) others. Researchers have referred to these techniques as "sleeper agents" since their effect can go unnoticed until invoked at a particular time or by a specific directive.[58] Research into the influence AI systems can have on user opinions validated the power of AI censorship. A small prompt injection, undetectable by users, significantly altered user opinions, illustrating a simple but effective model diversion method.[59]

A notable example of how building blocks enable censorship can be seen in word embedding. Word embedding is a key component of generative AI.[60] Algorithms like Word2Vec and GloVe map words into a vector space to mathematically reflect concept similarity. Typically, these word embeddings are trained on large, diverse datasets such as Wikipedia. Files of resulting pre-trained word vectors can be obtained for use in other natural language processing (NLP) applications including generative AI. Facebook provides such vector files for 294 different languages. Importantly, these files carry with them both semantic concepts and implicit biases.[61]

Research using Word2Vec to create word embedding vectors from both Baidu Baike, an online Chinese encyclopedia subject to government censorship, and Chinese language Wikipedia, free from government intervention, resulted in significantly different semantic connections. The authors note that "Chinese word embeddings trained with the same method but separately on these two corpuses reflect the political censorship of these corpuses, treating the concepts of democracy, freedom, collective action, equality, people and historical events in China significantly differently."[62]

Since these embedding files are utilized in a variety of other systems, the implicit bias has critical "downstream effects" acting as "force multipliers for censorship."[63] These censorship elements can be embedded unknowingly (or knowingly) in downstream systems, a form of "model poisoning," perpetuating their impact on a larger scale.[64]

Some of these allegations of censorship can be attributed to the use of reinforcement learning human feedback (RLHF). Most production LLM's utilize this process as a means to fine-tune or adjust the baseline responses of the system. RLHF is essentially a guardrail tool that "involves a human stepping in to teach a model which responses are good, and which responses are bad" where "good" and "bad" typically align with community standards.[65] However, institutional, corporate or state actors can utilize RLHF to fine-tune the system to promote certain perspectives and preclude others, effectively using the technique as a tool for censorship. Recent developments with "reinforcement learning AI feedback" (RLAIF) replace human intervention with a machine learning model making this an "end-to-end" AI process ideal for AI censorship.[66]

f) Regulation and Legislation

Regulatory environments and legislative programs motivated by political objectives can be an explicit tool of AI censorship. Government-mandated censorship varies across the globe with authoritarian countries such as China and Russia exerting enormous control over what content is allowed online. In accordance with the "Gay Propaganda Law," Roskomnadzor, Russia's federal censorship agency, prohibits platforms from publishing material promoting non-traditional sexual relations and so-called LGBT "propaganda" effectively censoring marginalized voices. Technology companies are forced to comply or face substantial fines resulting in censorship in AI systems.

However, more indirect methods of suppression arising from regulations can also be effective. Executive Orders from the US President regarding DEI programs (diversity, equity, inclusion) (here and here) have resulted in extensive data (websites, documents, reports, even mentions) being removed from government sites and databases. While specifically targeting the public sector, they also have implications for the private sector.[67] In response, the Center for Disease Control (CDC) has been instructed to remove, withdraw, and edit scientific papers that mention or reference "pregnant people, transgender, binary, non-binary, gender, assigned at birth, binary [*sic*], non-binary [*sic*], cisgender, queer, gender identity, gender minority, anything with pronouns."[68]

While this is a traditional, and blatant, case of censorship it also results in AI censorship. CDC authored papers, now under revision orders, are posted on their website and made available through the National Library of Medicine's PubMed. Since both are important sites for training and fine-tuning data, the resulting AI models will now be censored. Similarly, the initiative from OpenAI to create a ChatGPT Gov for US government employees will be fine-tuned on censored training data. Since the CDC decision, many other government agencies have been subject to similar orders and requirements, resulting in extensive civil society efforts to preserve this data.[69]

g) Open Source Models and AI Censorship

An area of considerable debate in the AI community, and one with important implications for AI censorship, is open-source models. Many proprietary GPTs are built on widely available pre-trained, open-source models. These models are an important part of the AI landscape because they allow for modifications and adjustments unavailable to users of commercially provided systems (e.g. ChatGPT, Gemini, etc.).[70] This is especially important for those wanting to fine-tune the pre-trained systems with their own local data where access to the architecture and hyperparameters are necessary. However, this level of access and modification makes it far easier for malicious action.[71] Open-source models are the core of many of the AI systems with specific censorship objectives. These models are accessible, customizable, and easily fine-tuned to suppress information and flood the system with other viewpoints. Sophisticated techniques such as privileged instructions and safety classifiers can be implemented in the models and used, not as guardrails, but rather as sleeper agents, directing and diverting output.[72]

Frameworks for AI Literacy

Duri Long and Brian Magerko define AI literacy as "a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace."[73] There are a number of valuable frameworks for AI literacy. Some come from an LIS perspective.[74] Other. frameworks represent other disciplinary approaches.[75] Still others define the competencies that form the desired outcomes of AI literary initiatives.[76] A

valuable analysis that sought to identify the determinants of AI literacy focused on the digital divide (access to technology), cognitive absorption (level of engagement with technology), and computational thinking (problem solving and conceptual skills).[77] All these AI literacy frameworks, competencies, and determinants identified issues regarding bias and misinformation, and underscored the importance of critical thinking in assessing the outcome of AI tools and services. However, none identified AI censorship as an element of AI literacy or an issue to be addressed.

Distinctive among these is the UNESCO AI competency framework for students. This report advises that students must "understand that all technical systems are socio-technical systems, and that socio-technical systems serve political agendas and are not neutral sources of information. Students engage concepts such as the stated and hidden goals of algorithms, algorithmic bias and human agency."[78] By identifying the political motivations guiding some (if not all) AI systems, the UNESCO report opens the door for discussions about AI censorship. Samuel Woolley and Philip Howard call this "computational propaganda," describing it as "interactive and ideological imbued [and] almost pure examples of politics in code" forming "the latest, and most ubiquitous, technical strategies to be deployed by those who wish to use information technology for social control."[79]

Enhancing AI Literacy: The Political Dimension

As Christine Jenkins notes "all censorship is political, with 'political' being not simply partisan politics, but rather the strategies used in claiming, wielding and defending social power, and thus controlling human societies."[80] To address AI censorship, it is necessary to focus on the political dimensions of AI literacy. However, as Annemaree Lloyd notes, the challenge of AI literacy is a "wicked problem" because

> algorithmic literacy differs from digital literacy, which focuses on core information literacy skills in the digital context, because it requires examination of culture (in both analogue and digital spaces), as a generative proposition and the construction of algorithms should be viewed as a practice which influences other aspects of social life.[81]

Addressing the political in the context of AI literacy, acknowledges that AI is not neutral.[82] These systems are

> designed to invite and shape participation, toward particular ends. This includes what kind of participation they invite and encourage; what gets displayed first or most prominently; how the platforms design navigation from content to user to exchange; the pressures exerted by pricing and revenue models; and how they organize information through algorithmic sorting, privileging some content over others, in opaque ways.[83]

As a result, it is necessary to adopt a critical focus. Critical AI literacy is "the ability to comprehend the core features of an AI system and its (in-)compatibility with its particular application contexts in a (necessarily) more complex sociotechnical reality."[84] To illustrate these, two perspectives regarding AI censorship and the relation to AI

literacy will be discussed: 1) the personal: AI developers and the AI lifecycle, and 2) the corporate: owners and leadership, and AI system policies.

a) The Personal

Human choices and biases are reflected in every decision in the AI development lifecycle, including “data collection and selection; data annotation; model development and evaluation; and model deployment, monitoring, and maintenance.”[85] As Jenna Burrell and Marion Fourcade observe, “far from being purely mechanistic, it [AI] is deeply, inescapably human.”[86] At each stage of the lifecycle, the biases of AI developers may lead to censorship, whether implicit or explicit, favouring dominant views in society while marginalizing or excluding other viewpoints.

For example, Will Orr notes that “datasets are products of their sociotechnical contexts” and that “data cleaning is not a value-neutral practice, and the decisions made during its creation mold the political implications of a dataset and the voices represented within it.”[87] By applying a critical lens, users can more thoroughly evaluate where the data originated, how the data were generated, the level of detail and granularity, and how the data was processed.[88] Data that are not representative, taken from sources that are restricted or censored, or labeled in a way that reproduces societal biases can lead to the reinforce of existing power imbalances in AI models effectively introducing censorship.

b) The Corporate

Corporations are by their nature political entities. As a result, the ownership and leadership of companies involved in AI systems must also be closely examined using a critical approach. This requires examining the power dynamics, economic motivations, beliefs, values, and ideology of the owners of technology companies to understand how this impacts the ways that social groups are, or are not represented, whose ideologies get supported, and whether technology pushes back against or upholds existing social hierarchies. The personal ideology of platform owners can shape policies such as content moderation, type of business model, corporate culture, and staffing decisions.

For example, when Elon Musk acquired Twitter (X), he chose to take the company private, allowing him to reinstate accounts that violated Twitter’s policies, roll back misinformation policies, reduce or disband trust and safety teams, loosen content moderation practices, and end access to data for researchers, all with little oversight.[89] However, as Adrian Kopps notes,

> policies represent only one aspect of X’s transformation. Changes to the platform’s functionalities, algorithmic operations and overall culture are other important factors to consider. Musk’s very outspoken political agenda and his self-proclaimed crusade against the “woke mind virus”, raises concerns about how the richest man in the world is using his power to influence which voices are suppressed and amplified on the platform.[90]

Conclusion

The pernicious effects of censorship are persistent and debilitating. As Nadine Gordimer, South African novelist and Nobel winner notes, “censorship is never over for those who have experienced it. It is a brand on the imagination that affects the individual who has suffered it, forever.”[91] Libraries and librarians have long understood this and have made anti-censorship efforts central to their values, policies, and practices. AI censorship presents a new challenge.

While there are grave concerns about misinformation, disinformation, deepfakes, bias, discrimination, and cybersecurity caused by malicious uses of AI, censorship by AI has only recently received warranted attention. Addressing this issue is urgent because employing AI techniques results in censorship that is “more precise [and] less detectable.”[92] However, unlike concerns about algorithmic recommendations or misinformation, which form the basis of much AI literacy programming, AI censorship awareness and detection asks not “what am I seeing and why” but rather “what I am *not* seeing and why?”

Responding to AI censorship through AI literacy requires enhancing existing approaches. As the UNESCO report highlights, it is the “political” dimension of AI systems that requires emphasis and scrutiny. The controlling interests of those providing these tools (whether state or corporate) must be recognized, understood, and mitigated as necessary. Libraries can respond to this new challenge by augmenting their AI literacy programming to identify censorship as an aspect of critical information literacy, to include a political analysis of AI, and to enhance awareness of the techniques of AI censorship that might impact information sources and tools.

Endnotes

[1] Jennifer Elaine Steele, “A History of Censorship in the United States,” *Journal of Intellectual Freedom & Privacy* 5, no. 1 (2020): 16, https://doi.org/10.5860/jifp.v5i1.7208.

[2] Cara S. Bertram, “Censorship Throughout the Centuries,” *American Libraries*, 2024; Elizabeth A. Harris, “Removing Books from Libraries Often Takes Debate. But There’s a Quieter Way.,” *The New York Times*, October 8, 2024, sec. Books, https://www.nytimes.com/2024/10/08/books/book-ban-library-weeding.html; Amanda Jones, *That Librarian: The Fight Against Book Banning in America* (New York: Bloomsbury, 2024).

[3] Philip Seargeant, “How AI Threatens Free Speech – and What Must Be Done about It,” The Conversation, January 18, 2024, http://theconversation.com/how-ai-threatens-free-speech-and-what-must-be-done-about-it-221330.

[4] Gary Marchionini, “Information and Library Professionals’ Roles and Responsibilities in an AI-Augmented World,” *Journal of the Association for Information Science and Technology* 75, no. 8 (2024): 865–68, https://doi.org/10.1002/asi.24930; Adrian Shahbaz, Allie Funk, and Kian Vesteinsson, “Freedom on the Net 2023: The Repressive Power of Artificial Intelligence” (Freedom House, 2023), https://freedomhouse.org/sites/default/files/2023-11/FOTN2023Final.pdf.

[5] Rishi Bommasani et al., “On the Opportunities and Risks of Foundation Models,” *arXiv*, 2021, http://arxiv.org/abs/2108.07258; Peter Slattery et al., “The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence,” 2024, https://doi.org/10.13140/RG.2.2.28850.00968; Laura Weidinger et al., “Taxonomy of Risks Posed by Language Models,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22 (New York, NY, USA: Association for Computing Machinery, 2022), 214–29, https://doi.org/10.1145/3531146.3533088.

[6] Gary Marcus, *Taming Silicon Valley: How We Can Ensure That AI Works for Us* (Cambridge MA: MIT Press, 2024).

[7] “ARL/CNI AI Scenarios: AI-Influenced Futures” (Washington, DC, and West Chester, PA: Association of Research Libraries, Coalition for Networked Information, and Stratus Inc, 2024), https://doi.org/10.29242/report.aiscenarios2024.; Ryan Cordell, “Machine Learning + Libraries: A Report on the State of the Field” (Washington DC: Library of Congress, 2020), https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf; Ryan Cordell, “Closing the Loop: Bridging Machine Learning (ML) Research and Library Systems,” *Library Trends* 71, no. 1 (2023): 132–43, https://doi.org/10.1353/lib.2023.0008; Andrea Gasparini and Heli Kautonen, “Understanding Artificial Intelligence in Research Libraries: An Extensive Literature Review,” *LIBER Quarterly: The Journal of the Association of European Research Libraries* 32, no. 1 (2022), https://doi.org/10.53377/lq.10934.

[8] Amy B. James and Ellen Hampton Filgo, “Where Does ChatGPT Fit into the Framework for Information Literacy? The Possibilities and Problems of AI in Library Instruction,” *College & Research Libraries News*

84, no. 9 (2023): 342, https://doi.org/10.5860/crln.84.9.334; Miriam W. Ndungu, "Integrating Basic Artificial Intelligence Literacy into Media and Information Literacy Programs in Higher Education: A Framework for Librarians and Educators," *Journal of Information Literacy* 18, no. 2 (2024): 122–39, http://dx.doi.org/10.11645/18.2.641.

[9] Slattery et al., "The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence," 36.

[10] American Library Association, "Challenge Support," 2022, https://www.ala.org/tools/challengesupport.

[11] Canadian Federation of Library Associations, "Statement on Intellectual Freedom and Libraries," 2019, https://cfla-fcab.ca/en/guidelines-and-position-papers/statement-on-intellectual-freedom-and-libraries/.

[12] Summer Lopez, "Speech in the Machine: Generative AI's Implications for Free Expression" (PEN America, 2023), https://pen.org/report/speech-in-the-machine/.

[13] Steven Feldstein, "The Road to Digital Unfreedom: How Artificial Intelligence Is Reshaping Repression," *Journal of Democracy* 30, no. 1 (2019): 51, https://dx.doi.org/10.1353/jod.2019.0003.

[14] H. Akin Ünver, "Artificial Intelligence, Authoritarianism and the Future of Political Systems" (Istanbul: EDAM, 2018), 16, https://edam.org.tr/wp-content/uploads/2018/07/AKIN-Artificial-Intelligence_Bosch-3.pdf.

[15] Shahbaz, Funk, and Vesteinsson, "Freedom on the Net 2023: The Repressive Power of Artificial Intelligence," 13.

[16] Dan Hendrycks and Mantas Mazeika, "X-Risk Analysis for AI Research" (arXiv, September 20, 2022), 13, https://doi.org/10.48550/arXiv.2206.05862.

[17] Yuval Noah Harari, *Nexus: A Brief History of Information Networks from the Stone Age to AI* (Random House, 2024), 313.

[18] Omaima Almatrafi, Aditya Johri, and Hyuna Lee, "A Systematic Review of AI Literacy Conceptualization, Constructs, and Implementation and Sssessment Efforts (2019–2023)," *Computers and Education Open* 6 (2024), https://doi.org/10.1016/j.caeo.2024.100173; Marc Pinski and Alexander Benlian, "AI Literacy for Users – A Comprehensive Review and Future Research Directions of Learning Methods, Components, and Effects," *Computers in Human Behavior: Artificial Humans* 2, no. 1 (January 1, 2024): 100062, https://doi.org/10.1016/j.chbah.2024.100062.

[19] Leon Carnovsky, "The Obligations and Responsibilities of the Librarian Concerning Censorship," *The Library Quarterly* 20, no. 1 (1950): 21–32, https://doi.org/10.1086/617600.

[20] Marchionini, "Information and Library Professionals' Roles and Responsibilities in an AI-Augmented World," 867.

[21] Martin Frické, *Artificial Intelligence and Librarianship: Notes for Teaching*, 3rd ed. (SoftOption, 2024), 344, https://softoption.us/sites/default/files/AIinLibrariesNotesForTeaching.pdf.

[22] Louise Amoore, "Machine Learning Political Orders," *Review of International Studies* 49, no. 1 (2023): 20–36, https://doi.org/10.1017/S0260210522000031.

[23] Iryna Pentina, Tyler Hancock, and Tianling Xie, "Exploring Relationship Development with Social Chatbots: A Mixed-Method Study of Replika," *Computers in Human Behavior* 140 (2023): 1–15, https://doi.org/10.1016/j.chb.2022.107600.

[24] Raja Parasuraman and Victor Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* 39, no. 2 (1997): 230–53, https://doi.org/10.1518/001872097778543886.

[25] Sarah Zheng, "China's Answers to ChatGPT Have a Censorship Problem," *Bloomberg.Com*, May 2, 2023, https://www.bloomberg.com/news/newsletters/2023-05-02/china-s-chatgpt-answers-raise-questions-about-censoring-generative-ai.

[26] Allie Funk, Adrian Shahbaz, and Kian Vesteinsson, "AI Chatbots Are Learning to Spout Authoritarian Propaganda," Wired, 2023, https://www.wired.com/story/chatbot-censorship-china-freedom-house; Shahbaz, Funk, and Vesteinsson, "Freedom on the Net 2023: The Repressive Power of Artificial Intelligence"; Ksenia Yermoshina, "Ukraine, Protests and Sexism. Ksenia Yermoshina on Censorship in Neural Networks," *Teplitsa. Technologies for Social Good*, June 15, 2023, https://te-st.org/2023/06/15/ai-censorship.

[27] Zeyi Yang, "There's No Tiananmen Square in the New Chinese Image-Making AI," *MIT Technology Review* (blog), September 14, 2022, https://www.technologyreview.com/2022/09/14/1059481/baidu-chinese-image-ai-tiananmen.

[28] Yermoshina, "Ukraine, Protests and Sexism. Ksenia Yermoshina on Censorship in Neural Networks."

[29] Zheng, "China's Answers to ChatGPT Have a Censorship Problem."

[30] Luiza Jarovsky, “DeepSeek’s Legal Pitfalls,” *Emerging AI Governance Challenges* (blog), February 2, 2025, https://www.luizasnewsletter.com/p/deepseeks-legal-pitfalls; Vivian Wang, “How Does DeepSeek’s A.I. Chatbot Navigate China’s Censors? Awkwardly.,” *The New York Times*, January 29, 2025, sec. World, https://www.nytimes.com/2025/01/29/world/asia/deepseek-china-censorship.html.

[31] Charles Rollet, “Leaked Data Exposes a Chinese AI Censorship Machine,” *TechCrunch* (blog), March 26, 2025, https://techcrunch.com/2025/03/26/leaked-data-exposes-a-chinese-ai-censorship-machine.

[32] Andy Lee Roth, avram anderson, and Mickey Huff, “Beyond Prior Restraint: Censorship by Proxy and the New Digital Gatekeeping” (Project Censored, 2024), https://www.projectcensored.org/beyond-prior-restraint-censorship-by-proxy-and-the-new-digital-gatekeeping.

[33] Rachel Griffin, “The Politics of Algorithmic Censorship: Automated Moderation and Its Regulation,” in *Music and the Politics of Censorship: From the Fascist Era to the Digital Age*, ed. James Garratt (Brepols, 2025), https://sciencespo.hal.science/hal-04325979.

[34] “The Santa Clara Principles on Transparency and Accountability in Content Moderation,” 2021, https://santaclaraprinciples.org.

[35] Jennifer Cobbe, “Algorithmic Censorship by Social Platforms: Power and Resistance,” *Philosophy & Technology* 34, no. 4 (2021): 762, https://doi.org/10.1007/s13347-020-00429-0.

[36] Laura Weidinger et al., “Ethical and social risks of harm from language models” (DeepMind, 2021), https://deepmind.com/research/publications/2021/ethical-and-social-risks-of-harm-from-language-models.

[37] Haimson, Oliver L., Daniel Delmonaco, Peipei Nie, and Andrea Wegner. “Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas.” *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (2021): 446:1-466:35. https://doi.org/10.1145/3479610.

[38] Lopez, “Speech in the Machine: Generative AI’s Implications for Free Expression.”

[39] “The Santa Clara Principles on Transparency and Accountability in Content Moderation.”

[40] Shahbaz, Funk, and Vesteinsson, “Freedom on the Net 2023: The Repressive Power of Artificial Intelligence,” 14.

[41] David Rozado, “The Political Biases of ChatGPT,” *Social Sciences* 12, no. 3 (2023): 1–8, https://doi.org/10.3390/socsci12030148; Anthony Zappin et al., “YouTube Monetization and Censorship

by Proxy: A Machine Learning Prospective," *Procedia Computer Science*, 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, 198 (January 1, 2022): 23–32, https://doi.org/10.1016/j.procs.2021.12.207.

[42] Joel Finkelstein et al., "The CCP's Digital Charm Offensive: How TikTok's Search Algorithm and pro-China Influence Networks Indoctrinate GenZ Users in the United States" (Network Contagion Research Institute, 2024), 2, https://networkcontagion.us/wp-content/uploads/NCRI-Report_-The-CCPs-Digital-Charm-Offensive.pdf.

[43] Committee on the Judiciary, and the, and Select Subcommittee on the Weaponization of the Federal Government, "The Weaponization of the National Science Foundation: How NSF Is Funding the Development of Automated Tools to Censor Online Speech 'at Scale' and Trying to Cover up Its Actions" (Washington D.C.: U.S. House of Representatives, 2024), https://judiciary.house.gov/sites/evo-subsites/republicans-judiciary.house.gov/files/evo-media-document/NSF-Staff-Report_Appendix.pdf.

[44] Mark Zukerberg, "It's Time to Get Back to Our Roots Around Free Expression [Video]," Facebook, January 7, 2025, https://www.facebook.com/zuck/videos/1525382954801931.

[45] Stuart A. Thompson, "Uncensored Chatbots Provoke a Fracas Over Free Speech," *The New York Times*, July 2, 2023, sec. Technology, https://www.nytimes.com/2023/07/02/technology/ai-chatbots-misinformation-free-speech.html; Maxwell Zeff and Thomas Germain, "We Tested AI Censorship: Here's What Chatbots Won't Tell You," Gizmodo, March 29, 2024, https://gizmodo.com/we-tested-ai-censorship-here-s-what-chatbots-won-t-tel-1851370840.

[46] Pranav Dixit, "This Uncensored Chatbot Shows What Happens When AI Is Programmed to Disregard Human Decency," BuzzFeed News, March 29, 2023, https://www.buzzfeednews.com/article/pranavdixit/freedomgpt-ai-chatbot-test.

[47] Stefan Baack, "A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA: ACM, 2024), 2199, https://doi.org/10.1145/3630106.3659033.

[48] Common Crawl, "Common Crawl - Open Repository of Web Crawl Data," 2024, https://commoncrawl.org.

[49] Mohamed Ahmed and Jeffrey Knockel, "The Impact of Online Censorship on LLMs," *Free and Open Communications on the Internet*, no. 2 (2024), https://www.petsymposium.org/foci/2024/foci-2024-0006.pdf.

[50] Ahmed and Knockel, 2.

[51] Eddie Yang and Margaret E. Roberts, "The Authoritarian Data Problem," *Journal of Democracy* 34, no. 4 (2023): 141–50.

[52] Baack, "A Critical Analysis of the Largest Source for Generative AI Training Data."

[53] Baack; Ana-Andreea Stoica, Nelly Litvak, and Augustin Chaintreau, "Fairness Rising from the Ranks: HITS and PageRank on Homophilic Networks," in *Proceedings of the ACM Web Conference 2024*, WWW '24 (New York, NY, USA: Association for Computing Machinery, 2024), 2594–2602, https://doi.org/10.1145/3589334.3645609.

[54] Baack, "A Critical Analysis of the Largest Source for Generative AI Training Data," 2204.

[55] Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, edited by Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, 1286–1305. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.98.

[56] Remi Denton et al., "Bringing the People Back In: Contesting Benchmark Machine Learning Datasets" (arXiv, 2020), https://doi.org/10.48550/arXiv.2007.07399.

[57] Will Orr and Kate Crawford, "The Social Construction of Datasets: On the Practices, Processes, and Challenges of Dataset Creation for Machine Learning," *New Media & Society* 26, no. 9 (2024): 4955, https://doi.org/10.1177/14614448241251797.

[58] Evan Hubinger et al., "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training" (arXiv, 2024), https://doi.org/10.48550/arXiv.2401.05566.

[59] Maurice Jakesch et al., "Co-Writing with Opinionated Language Models Affects Users' Views," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23 (New York, NY, USA: Association for Computing Machinery, 2023), 1–15, https://doi.org/10.1145/3544548.3581196.

[60] Deepak Suresh Asudani, Naresh Kumar Nagwani, and Pradeep Singh, "Impact of Word Embedding Models on Text Analytics in Deep Learning Environment: A Review," *Artificial Intelligence Review*, 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare, 56, no. 9 (2023): 10345–425, https://doi.org/10.1007/s10462-023-10419-1.

[61] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-like Biases," *Science* 356, no. 6334 (2017): 183–86, https://doi.org/10.1126/science.aal4230.

[62] Eddie Yang and Margaret E. Roberts, "Censorship of Online Encyclopedias: Implications for NLP Models," in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, ACM, 2021, 537, https://doi.org/10.1145/3442188.3445916.

[63] Shahbaz, Funk, and Vesteinsson, "Freedom on the Net 2023: The Repressive Power of Artificial Intelligence," 13; Yang and Roberts, "Censorship of Online Encyclopedias."

[64] Hubinger et al., "Sleeper Agents."

[65] Zeff and Germain, "We Tested AI Censorship."

[66] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback" (arXiv, December 15, 2022), https://doi.org/10.48550/arXiv.2212.08073.

[67] Julia E. Judish et al., "Trump Executive Orders Target DEI in Government and Private Sector," *Pillsbury Law* (blog), January 28, 2025, https://www.pillsburylaw.com/en/news-and-insights/trump-anti-dei-executive-orders.html.

[68] Katherine J. Wu, "CDC Data Are Disappearing," *The Atlantic* (blog), January 31, 2025, https://www.theatlantic.com/health/archive/2025/01/cdc-dei-scientific-data/681531.

[69] Julian Lucas, "The Volunteer Data Hoarders Resisting Trump's Purge," *The New Yorker*, March 14, 2025, https://www.newyorker.com/news/the-lede/the-data-hoarders-resisting-trumps-purge.

[70] Ben Brooks, "Open-Source AI Is Good for Us," *IEEE Spectrum*, February 8, 2024, https://spectrum.ieee.org/open-source-ai-good.

[71] David Evan Harris, "Open-Source AI Is Uniquely Dangerous," *IEEE Specturm*, January 12, 2024, https://spectrum.ieee.org/open-source-ai-2666932122.

[72] Jinhwa Kim, Ali Derakhshan, and Ian G. Harris, “Robust Safety Classifier for Large Language Models: Adversarial Prompt Shield” (arXiv, October 31, 2023), https://doi.org/10.48550/arXiv.2311.00172; Eric Wallace et al., “The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions” (arXiv, 2024), https://doi.org/10.48550/arXiv.2404.13208.

[73] Duri Long and Brian Magerko, “What Is AI Literacy? Competencies and Design Considerations,” in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI ’20 (Honolulu, HI, USA: Association for Computing Machinery, 2020), 2, https://doi.org/10.1145/3313831.3376727.

[74] Sandy Hervieux and Amanda Wheatley, “Building an AI Literacy Framework: Perspectives from Instruction Librarians and Currently Information Literacy Tools” (Choice, 2024), https://www.choice360.org/research/white-paper-building-an-ai-literacy-framework-perspectives-from-instruction-librarians-and-current-information-literacy-tools; Ali Shiri, “Artificial Intelligence Literacy: A Proposed Faceted Taxonomy,” *Digital Library Perspectives* (January 1, 2024), https://doi.org/10.1108/DLP-04-2024-0067.

[75] Melanie Hibbert, Elana Altman, and Tristan Shippen, “A Framework for AI Literacy,” *EDUCAUSE Review: Emerging Technologies and Trends* (blog), June 3, 2024, https://er.educause.edu/articles/2024/6/a-framework-for-ai-literacy; Kathryn MacCallum, David Parsons, and Mahsa Mohaghegh, “The Scaffolded AI Literacy (SAIL) Framework for Education: Preparing Learners at All Levels to Engage Constructively withArtificial Intelligence,” *Hi Rourou* 1, no. 1 (2024), https://doi.org/10.54474/herourou.v1i1; Peter Tiernan et al., “Information and Media Literacy in the Age of AI: Options for the Future,” *Education Sciences* 13, no. 9 (2023), https://doi.org/10.3390/educsci13090906.

[76] Ravinithesh Annapureddy, Alessandro Fornaroli, and Daniel Gatica-Perez, “Generative AI Literacy: Twelve Defining Competencies,” *Digital Government: Research and Practice*, 2024, https://doi.org/10.1145/3685680; Divina Frau-Meigs, “Algorithm Literacy as a Subset of Media and Information Literacy: Competences and Design Considerations,” *Digital* 4, no. 2 (2024): 512–28, https://doi.org/10.3390/digital4020026.

[77] Ismail Celik, “Exploring the Determinants of Artificial Intelligence (AI) Literacy: Digital Divide, Computational Thinking, Cognitive Absorption,” *Telematics and Informatics* 83 (2023), https://doi.org/10.1016/j.tele.2023.102026.

[78] Fengschun Miao and Kellhy Shiohira, “AI Competency Framework for Students” (UNESCO, 2024), 66, https://unesdoc.unesco.org/ark:/48223/pf0000391105.locale=en.

[79] Samuel C. Woolley and Philip N. Howard, “Automation, Algorithms, and Politics | Political Communication, Computational Propaganda, and Autonomous Agents,” *International Journal of Communication* 10 (2016): 4886.

[80] Andrea Lynn, “You Mean People Still Try to Ban Books They Don’t Like?!,” *New Bureau. University of Illinois Urbana-Champaign*, September 22, 2006, https://news.illinois.edu/you-mean-people-still-try-to-ban-books-they-dont-like.

[81] Annemaree Lloyd, “Chasing Frankenstein’s Monster: Information Literacy in the Black Box Society,” *Journal of Documentation* 75, no. 6 (2019): 1483, https://doi.org/10.1108/JD-02-2019-0035.

[82] Annie Pho and Wynn Tranfield, “Building the Path for the Last Mile: Developing Critical AI Literacy for Library Workers,” *Journal of Radical Librarianship* 10 (2024): 178–93. https://journal.radicallibrarianship.org/index.php/journal/article/view/112.

[83] Tarleton Gillespie, “Regulation of and by Platforms,” in *The SAGE Handbook of Social Media*, ed. Jean Burgess, Alice Marwick, and Thomas Poell (Sage Publications, 2018), 257.

[84] Stefan Strauß, “Don’t Let Me Be Misunderstood: Critical AI Literacy for the Constructive Use of AI Technology,” *TATuP - Zeitschrift Für Technikfolgenabschätzung in Theorie Und Praxis* 30, no. 3 (2021): 45, https://doi.org/10.14512/tatup.30.3.44.

[85] You Chen et al., “Human-Centered Design to Address Biases in Artificial Intelligence,” *Journal of Medical Internet Research* 25, no. 1 (2023): 3, https://doi.org/10.2196/43251.

[86] Jenna Burrell and Marion Fourcade, “The Society of Algorithms,” *Annual Review of Sociology* 47, no. 1 (2021): 231, https://doi.org/10.1146/annurev-soc-090820-020800.

[87] Will Orr, “9 Ways to See A Dataset: Datasets as Sociotechnical Artifacts — The Case of ‘Colossal Cleaned Common Crawl’ (C4),” Knowing Machines, 2023, https://knowingmachines.org/publications/9-ways-to-see/essays/c4.

[88] Alan Freihof Tygel and Rosana Kirsch, “Contributions of Paulo Freire for a Critical Data Literacy: A Popular Education Approach,” *The Journal of Community Informatics* 12, no. 3 (2016): 108–21, https://doi.org/10.15353/joci.v12i3.3279.

[89] Elizabeth Blakey, “The Day Data Transparency Died: How Twitter/X Cut Off Access for Social Research,” *Contexts* 23, no. 2 (2024): 30–35, https://doi.org/10.1177/15365042241252125.

[90] Adrian Kopps, “Two Years After the Takeover: Four Key Policy Changes of X Under Musk” (Zenodo, 2024), 5, https://doi.org/10.5281/zenodo.14040407.

[91] Gordimer, Nadine. “Censorship and Its Aftermath.” *Index of Censorship* 7 (1990): 14–16.

[92] Shahbaz, Funk, and Vesteinsson, “Freedom on the Net 2023: The Repressive Power of Artificial Intelligence,” 13.