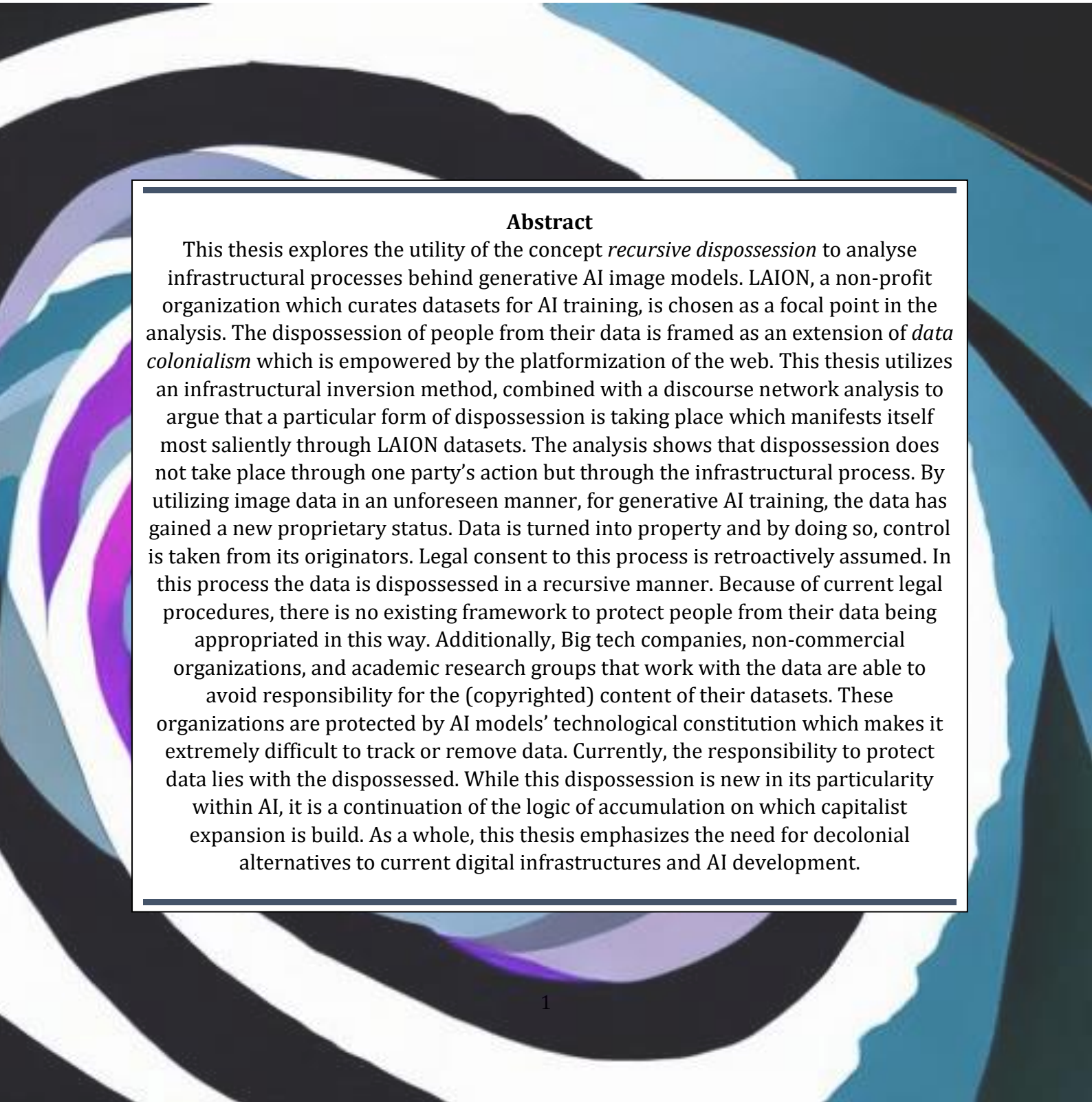
MA New Media & Digital Culture

LAION: Image Data, AI, and Dispossession

Recursive dispossession of image data as the result of Open Web infrastructure for AI generative models' development

Laura Jannes Burger
6380816



Abstract

This thesis explores the utility of the concept *recursive dispossession* to analyse infrastructural processes behind generative AI image models. LAION, a non-profit organization which curates datasets for AI training, is chosen as a focal point in the analysis. The dispossession of people from their data is framed as an extension of *data colonialism* which is empowered by the platformization of the web. This thesis utilizes an infrastructural inversion method, combined with a discourse network analysis to argue that a particular form of dispossession is taking place which manifests itself most saliently through LAION datasets. The analysis shows that dispossession does not take place through one party's action but through the infrastructural process. By utilizing image data in an unforeseen manner, for generative AI training, the data has gained a new proprietary status. Data is turned into property and by doing so, control is taken from its originators. Legal consent to this process is retroactively assumed. In this process the data is dispossessed in a recursive manner. Because of current legal procedures, there is no existing framework to protect people from their data being appropriated in this way. Additionally, Big tech companies, non-commercial organizations, and academic research groups that work with the data are able to avoid responsibility for the (copyrighted) content of their datasets. These organizations are protected by AI models' technological constitution which makes it extremely difficult to track or remove data. Currently, the responsibility to protect data lies with the dispossessed. While this dispossession is new in its particularity within AI, it is a continuation of the logic of accumulation on which capitalist expansion is build. As a whole, this thesis emphasizes the need for decolonial alternatives to current digital infrastructures and AI development.

Master's thesis for New Media & Digital Culture

Title: *LAION: Image Data, AI, and Dispossession*

Name: Laura Jannes Burger

Student Number: 6380816

Thesis Supervisor: Gerwin van Schie

Date: 01/05/2023

Words: 12741

Cover image made with Stable Diffusion

Contents

§1. Introduction (1.200 words)

Artificial intelligence art generators as a digital tool have become accessible to internet users in the past two years. While some people have hailed the tools as a new era for artistry, Others have not been quite positive.[1] They have taken issue with their images being used to train the tools without their knowledge or consent.[2] Especially professional graphic artists regard their art being used to train these generative models as theft.[3] As a response artists and legal scholars have started calling for stricter copyright laws to combat this, but current legal frameworks are so far insufficient.[4] One artist aptly described generative art models as "washing machines of intellectual property."[5] Artificial intelligence programs that utilize data from the Web are becoming more common and have a wide arrange of purposes. Third-party companies are using the accessible data for their own programs (e.g., Clearview and ChatGPT), raising privacy concerns and ethical dilemmas.[6] Are these the growing pains of a new technology, that we will find frivolous concerns in a couple of years, or is there another frame through which we can already understand what is happening? I argue that the web-based *infrastructure* which is supporting many types of AI development is this washing machine as it changes the dynamics of data as property. The central contribution of this thesis is in discerning how data is being appropriated through the lens of *recursive dispossession*, which furthers the observation that this 'washing machine' is not new, but rather a continuation of the capitalist logic of accumulation for profit. Approaching the topic from a postcolonial perspective as its knowledge provides a critical lens through which to analyse power structures and inequalities.

The focus of this analysis are the datasets made by LAION, as a centre-node in this infrastructure. When creating an AI art tool, it is trained on hundreds of millions of images before it produces passable images. LAION is a source for retrieving such large datasets. The accessibility of data makes it staggeringly easy to use and difficult to trace, even for its developers. The patterns people see in AI generated images might not reflect how the AI operates.[7] I will argue

[1] Kevin Kelly, "Picture Limitless Creativity at Your Fingertips," Wired.com, November 17, 2022, https://www.wired.com/story/picture-limitless-creativity-ai-image-generators/.
[2] Cloe Xiang, "AI Is Probably Using Your Images and It's Not Easy to Opt Out," Vice.com, September 26, 2022, https://www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out.
[3] Luke Plunkett, "AI Creating 'Art' Is An Ethical And Copyright Nightmare," Kotaku.com, August 25, 2022. https://kotaku.com/ai-art-dall-e-midjourney-stable-diffusion-copyright-1849388060.
[4] Celine Melanie A. Dee, "Examining Copyright Protection of AI-Generated Art," *Delphi - Interdisciplinary Review of Emerging Technologies* 1 (2018): 31-37.
[5] Plunkett, "AI Creating 'Art' Nightmare."
[6] Camilla Dul, "Facial Recognition Technology vs Privacy: The Case of Clearview AI," *Queen Mary Law Journal* (2022): 1-24, https://heinonline.org/HOL/P?h=hein.journals/qmlj2022&i=11.
[7] Janus Rose, "Why Does This Horrifying Woman Keep Appearing in AI-Generated Images?" Vice.com, September 7, 2022, https://www.vice.com/en/article/g5vjw3/why-does-this-horrifying-woman-keep-appearing-in-ai-generated-images.

that a new unforeseeable proprietary status is attached to web-sourced image data. Despite developers' insistence on the datasets and models having democratizing effects, open-access datasets and AI models are most profitable for large tech companies.[8] I will argue this change in proprietary status has taken place by analysing the data relations in the infrastructure around LAION. The meaning of online image data itself significantly changes as it is used as an indefinitely expanding source for profit. It causes its originators to lose control over data they previously did not need control over, simultaneously becoming owners of property and having it taken from them. Which is a particular form of dispossession, *recursive dispossession*. Data is subjected to logics that mirror the historical processes and effects of colonialism. The data relations in question are referred to as *data colonialism.*[9] Building on the notion that platforms have become infrastructuralized, this research approaches these issues as a result of a complex network in digital media.[10] *Infrastructure* is a historically grown relational state of (smooth) coordinated work consisting of individual activities, which rely on different classifications.[11] The data these tools are trained on has to originate somewhere. However, infrastructures tend to disappear into the background, making the results that these tools produce seem neutral or 'natural,' but they are the manifestations of a long process of negotiation, decision-making, and management of complex relationships.[12] Using *infrastructural inversion*, I investigate how this infrastructure model is negotiated and what actors hold power to mould the relations. This method lays bare the logics of data relations. Revealing the established processes is a step towards understanding how contemporary data practices have created these emergent models. Therefore, I ask the question: *How do LAION's datasets contribute to recursive dispossession through its infrastructural functionality and how does it reflect the wider logics of data colonialism?*

In the subsequent section, the contribution of this research to the academic fields of media studies and critical data studies will be explained. This is followed by a comprehensive overview of its theoretical foundations supporting the analysis (§2). The methodology will then be discussed, outlining why the research utilises an infrastructural inversion and discourse network analysis, what this entails, and how the study will proceed (§3). The analysis itself has been

[8] Peter Cihon, Jonas Schuett, and Seth D. Baum, "Corporate Governance of Artificial Intelligence in the Public Interest," *Information* no. 12, 275 (2021): 1 – 2, https://doi.org/10.3390/info12070275.
[9] Nick Couldry, and Ulises A Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media* 20, no. 4 (2019): 336 – 349, https://doi.org/10.1177/1527476418796632.
[10] Jean-Christophe Plantin, Carl Lagoze, Paul N. Edwards, and Christian Sandvig, "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook," *New Media & Society* 20, no. 1 (2018): 308. https://doi.org/10.1177/1461444816661553.
[11] Wolfgang Kaltenbrunner, "Infrastructural Inversion As a Generative Resource in Digital Scholarship," *Science As Culture* 24, no. 1 (2015): 4. https://doi.org/10.1080/09505431.2014.917621.
[12] Bowker, Geoffrey C, and Susan Leigh Star. *Sorting Things Out: Classification and Its Consequences*. Inside Technology. Cambridge, Mass.: MIT Press, 1999.

divided in six parts in order to address the research question. I formulated a series of smaller questions to construct a comprehensive answer. The first question considers the naturalization of data and its subjection to colonial logics and its relation to dispossession (§4.1 & §4.2). By examining digital media in the context of data as a commodified resource, it was possible to explore the role of programs such as LAION in shaping these dynamics. Approaching data from this perspective allowed for a critical examination of the underlying infrastructure, revealing the assumptions around the nature of data that are often taken for granted. The second question explores the functions of LAION and the classification practices upon which it is based. How property emerged out of non-property. It is important to recognize that LAION did not arise out of thin air, and that the choices made under data colonialism are reflected in the database itself and in the culture it is a part of (§4.3 & §4.4). The third question is about how data and AI models are handled by legal, financial, and academic stakeholders. The inversion method was used to consider the political dimensions of LAION and to question who benefits from the power asymmetry that is proposed by data colonialism. Highlighting the processes and practical consequences of LAION will provide insight in who has power and how the data is seamlessly dispossessed from its originators (§4.5 & §4.6).

§1.1 Academic Embedding

By researching this database infrastructurally it considers the specific ways dispossession is manifesting through infrastructure. By applying Couldry and Mejias' assertions on data colonialism together with the infrastructuralization of platforms, the analysis shows the perils of platform ubiquity and oligarchic lead on data relations. This research builds on Couldry and Mejias' judgement that colonial data relations dispossess people's life in all its tenets. Commodifying human social relations, transforming everyday life into 'capitalist production' which is in the process of manifesting itself properly concerningly fast. They call for research on how this is happening in practice.[13] It also connects to Critical Data Studies in the investigations of data assemblages as complex sociotechnical systems that produce, manage, analyse and translate data for a particular purpose.[14] The infrastructure under analysis can be regarded as one such assemblage. The research contributes to the field of critical data studies by exploring one configuration of data processing. My assertion is that the issues around AI art tools is a manifestation of dispossession, and it simultaneously illustrates how colonial data relations currently operate with the infrastructure of digital media. As such it reasserts the urgency of

[13] Couldry and Mejias, "Data Colonialism," 343.
[14] Rob Kitchin, "The Data Revolution: A critical analysis of big data, open data and data infrastructures," *The Data Revolution* (2021): 22.

widely available and decolonial digital infrastructure as a foundation for social justice, as Plantin et al. stated.[15] The AI art is characteristic of the wider operating logics of infrastructuralized platforms. Despite this analysis sole focus on the image database LAION, it reflects the issues under study more broadly.

§2. Theoretical Framework (2.550 words)

Before contextualizing LAION within the platform infrastructure and understanding its significance, it is necessary to examine several theoretical concepts. The *platformization* of the internet has significant implications for digital infrastructures in two respects.[16] Firstly, the concentration of image data, as the majority of scraped data originates from a small number of domains. Secondly, the concentration of power. Large platforms not only earn immense profit by selling user data, but their services have also become essential to daily life. In addition, these platforms tend to invest in expanding technical capabilities which increases power asymmetries.[17] These effects of *platformization* are imperative to take into account when discussing the 'Open Web,' since these companies have influence in ways which are not always obvious. Next, I will discuss data, and in particular the datafied subject under contemporary capitalism. I will introduce the concept of *surveillance capitalism* which is the result of commodifying social relations into monetizable data.[18] Zuboff formulated the original theory about this process. Couldry and Mejias' postcolonial perspective is essential to create a comprehensive idea about the mechanics of capitalism, learning from similar phenomena in colonial history.[19] They describe current data relations as leading to a type of *dispossession*, namely of the self, due the ubiquitous digital surveillance.[20] However, this is not the only type of dispossession that data relations create. Here, postcolonial theory aids us once again, by relating it to the dispossession of land from the native peoples of America. This helps us to understand *recursive dispossession*: something unowned is made into property through the act of taking it.[21] The analysis, through careful examination of the infrastructure, I show that this type of dispossession provides a comprehensive frame to interpret the multiple issues around AI

[15] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 307.
[16] Anne Helmond, "The Platformization of the Web: Making Web Data Platform Ready." *Social Media Society* 1, no. 2 (2015): 1,https://doi.org/10.1177/2056305115603080.
[17] Cihon, Schuett, and Baum, "Corporate Governance of AI," 1 – 2.
[18] Shoshana Zuboff, "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization," *Journal of Information Technology* 30, no. 1 (2015): 75–89. https://doi.org/10.1057/jit.2015.5.
[19] Couldry and Mejias, "Data Colonialism."
[20] Couldry and Mejias, "Data Colonialism," 344.
[21] Robert Nichols, *Theft Is Property!: Dispossession and Critical Theory* (Duke University Press), 2020, 91, https://doi.org/10.2307/j.ctv11smqjz

generated images. The influence that resources platforms have, should not be underappreciated. As I am discussing a system that is not connected to a specific platform, I will argue that 'Big Data' companies will reap the largest benefit, as the recursive aspect of the dispossession legitimizes and solidifies the form of data relations that benefits them.

§2.1 The Platformization of the Open Web

Changes in the dominant infrastructures on the web in the past fifteen years are often characterized as a shift from a decentralized internet, where activities were performed on different domains, to a more centralized internet, where activities are performed on only a handful of platforms.[22] This shift had consequences for the power dynamics and data flows on the web. Platforms now provide vast numbers of services which are integral to the functioning of societies worldwide.[23] This was described eloquently by Anne Helmond in "The platformization of the web: Making web data platform ready," where she introduced the concept of *platformization*.[24] Through platformization data flows have become interchangeable, modulable, and centralized. Integrating separate corners of the web through API's (application programming interfaces) has been a deliberate project.[25] This process is crucial to the development of digital infrastructure and underlines its profit driven origins. Discussed here are the changes the web went through infrastructurally. *Infrastructure* is a historically grown relational state of (smooth) coordinated work consisting of individual activities, which rely on different classifications.[26] The possibility to platformize the web depended on the shifting standardization of web page coding. Initially, pages were coded in HTML which was legible enough for people but difficult for automated programs or bots to read, process, or reuse.[27] To facilitate the API projects, web pages started using standardized XML coding, which was made for machine extraction and with high cross compatibility.[28] Using social plug-ins, a form of API, they created a two-way data flow between websites and connecting platforms.[29] While the platform functionalities are decentralized, its data is recentralized to the platform. These developments set the stage for a handful of platforms to become behemoths that increasingly protrude into all creases off the internet.[30]

[22] Helmond, "Platformization," 1.
[23] José van Dijck, "Seeing the Forest for the Trees: Visualizing Platformization and Its Governance," *New Media & Society* 23, no. 9 (2021): 2808 – 2809, https://doi.org/10.1177/1461444820940293.
[24] Helmond, "Platformization."
[25] Helmond, "Platformization," 3.
[26] Wolfgang "Infrastructural Inversion," 4.
[27] Helmond, "Platformization," 3.
[28] Helmond, "Platformization," 5-6.
[29] Helmond, "Platformization," 6.
[30] Helmond, "Platformization," 1.

Plantin et al. call this the 'infrastructuralization of platforms' which is the increasing presence of digital media in everyday life as well as the integration of platforms within all web-based services.[31] Infrastructure studies tends to favour a historical perspective. Seeing the Open Web as an infrastructure that is constructed on principles made by people. Data flows which either originate as social media platforms capture vast amounts of data or originate through platform configurations are technically findable by anyone. People and organisations alike utilize platform services (like API) to save costs and increase cross-compatibility with their own websites. Meanwhile, data scrapers like Common Crawl make use of the Open Web infrastructure to collect data en masse, as much of the internet is accessible with few limits. This is coupled with the compounding programmability of its content, Platin et al. questions whether this infrastructure puts all this data in a precarious position for exploitation.[32] With this they mean that the 'openness' of the Open Web becomes problematic when it is interwoven in the platform power relations. Platin et al. call the Open Web, a publicly accessible architecture that is essentially a "global commons."[33] Social interactions are increasingly being performed on large platforms. As a result, these platforms hold an extreme amount of control over web-based services since user data are consistently profitable commodities. It might appear counter-intuitive to then claim that the openness of the web would embolden their reach. Since anyone can access the content on the internet, it should be democratizing, as many organizations including LAION itself claim.[34] Multiple studies have pointed out that this might be an utopic view on the matter that tends to brush of the nuances of collecting, managing, and responsibly storing 'freely available' data resources. Additional arguments are the lack of transparency - especially with respect to the 'right to be forgotten' - and the risk of repeating harmful biases.[35] Large platforms understand the potential financial value of any data type, so the current operationality of the Open Web results in an asymmetric and possibly exploitative infrastructure. This is why we should remain critical about whether the 'democratizing' argument holds water.[36] Therefore, we should consider what exploitation means here.

---

[31] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 294-295.
[32] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 302.
[33] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 302.
[34] Christoph Schuhmann et al., "Laion-5b: An open large-scale dataset for training next generation image-text models." *arXiv preprint arXiv:2210.08402* (October 16, 2022): 12, https://doi.org/10.48550/arXiv.2210.08402.
[35] Gertraud Koch and Katharina Kinder-Kurlanda, "Source Criticism of Data Platform Logics on the Internet," *Historical Social Research / Historische Sozialforschung 45*, no. 3 (2020): 271.
Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," *arXiv* (October 5, 2021): 1 -15.
Dul, "Facial Recognition Technology vs Privacy," 1-24.
[36] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 302.

§2.2 Data under Capitalism and Colonialism

The consolidation of power by platforms has allowed for new forms of capitalization, referred to as Surveillance Capitalism, a term that originated from Shoshana Zuboff's chapter "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization."[37] Zuboff argues that we are in the early stages of a new institutionalizing logic of accumulation that produces unfathomable amounts of data to monetize, control, and modify behaviour.[38] It considers the idea that data is being captured for ends that does not benefit humanity despite being rationalized as such. Before platformization, the internet was a decentralized and open place for an increasing amount of everyday activities to take place and to share quite intimate social moments as well. Tech companies, like Google, began accumulating large datasets based on the notion that it would be profitable, which they dubbed 'big data'.[39] They were correct and the 'Big Data' sector is tremendously profitable, creating a new architecture that supports this developing economy.[40] Under this logic, user's behaviour is modified by incentivizing or requiring people to engage in ways which produces the most (monetizable) data by probing into intimate and preferably private parts of their social life.[41] Additionally, their surveillance experiments and digital extraction are not processes societies have dealt with before. Thus, governments have no instruments to limit Google's activities which preserves the company's head start against actors upholding privacy principles.[42] Most actions to limit or regulate ethically have to be done retroactively, which becomes extremely problematic since platform services have become infrastructuralized. Zuboff poses that this dynamic between capital and people is now radically different since there is no exchange of labour and consumption, rather people are simply 'targets' for extraction. "Under surveillance capitalism, democracy no longer functions as a means to prosperity; democracy threatens surveillance revenues," implying a different dynamic to governance altogether.[43]

This different dynamic between capital and people can be understood through a postcolonial lens, as Couldry and Mejias do in their book *The Cost of Connection* where they frame the platform's profitmaking practices through the lens of historical colonial practices.[44] They

[37] Zuboff, "Surveillance Capitalism."
[38] Zuboff, "Surveillance Capitalism," 85.
[39] Jim Thatcher, David O'Sullivan, and Dillon Mahmoudi, "Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data," *Environment and Planning D: Society and Space* 34, no. 6 (2016): 992, https://doi.org/10.1177/0263775816633195.
[40] Zuboff, "Surveillance Capitalism," 78.
[41] Zuboff, "Surveillance Capitalism," 84.
[42] Zuboff, "Surveillance Capitalism," 83.
[43] Zuboff, "Surveillance Capitalism," 86.
[44] Nick Couldry and Ulises A. Mejias, *The costs of connection: How data is colonizing human life and appropriating it for capitalism* (Stanford University Press, 2020), 84.

draw parallels to colonialism's functioning as a process of forging new economies that spread incrementally.[45] Data itself is under scrutiny as they point to the configuration of data relations as a leading mechanism of that economic power. Data relations are commodified social relations. Arguing from a Marxist perspective, it is similar to how 'work' has been commodified into quantifiable labour, human social relations, expressed through social engagements on platforms, are commodified into quantifiable datapoints. Data relations are fundamentally human relations that can (eventually) be turned into monetizable datasets.[46] The role of platforms is to create the conditions under which people's social relations can be captured and appropriated into valuable commodities.[47] Due to the digital environment being constituted by complex computational models, extraction is difficult to fully comprehend for those unschooled on the subject. Combined with the dependence on web-based services for basic societal participation, people have been contributing to value extraction without reciprocity, thus subject to exploitation. This is "an appropriation that constitutes the colonial moment of contemporary capitalism."[48] There are certain rationalities and associated practices that allow people to build a system based on exploitation while obscuring or justifying said exploitation which mirrors historical colonialism. They specifically mention the data being positioned as 'open,' as in freely available for utilization. Similar to how 'natural resources' are set up as unproblematically extractable pieces of the world. These rationalities treat the foundational social act as valueless, 'just sharing,' naturalizing their extraction, raw materials only made into valuable products by companies.[49] Important to consider is that data appropriation is not a new form of labour, but rather another contributor to surplus value, sourced from social life. Like a stone quarry affects its surroundings by leaving a large hole, social relations will be affected by its appropriation as well. Social relations will incrementally be managed and reconfigured in service of value extraction.[50] While they say that "It is premature to map the forms of capitalism that will emerge from it on a global scale," the processes at work are incrementally seeping into everyday life.[51] Data relations are not only profiting from social relations in its new and specific manner, but are forging new social relations as well, identified as *data colonialism.* Couldry and Mejias and Zuboff point to ubiquitous surveillance, a panoptic force that digitally tracks and controls each terrain of human endeavour. What then is the cost? They consider this new power configuration a threat to the "bare reality of

[45] Couldry and Mejias, "Data Colonialism," 337.
[46] Couldry and Mejias, "Data Colonialism," 346.
[47] Couldry and Mejias, "Data Colonialism," 338.
[48] Couldry and Mejias, "Data Colonialism," 342.
[49] Couldry and Mejias, "Data Colonialism," 340.
[50] Couldry and Mejias, "Data Colonialism," 343.
[51] Couldry and Mejias, "Data Colonialism," 337.

the self," as data dispossesses the person from the self.[52] These lofty claims will be further discussed as we consider modes of *dispossession*.

§2.3 Recursive Dispossession

This is where the concept of dispossession aids us in seeing the process at play. At first, to frame dispossession as an ongoing process in digital infrastructures might appear far-fetched. In this point I believe Couldry and Mejias could have been more descriptive, as they pose the dispossession to be a replication of the extraction of natural resources in the colonial era.[53] Dispossession to them "happens through the appropriation of things that belong to someone else and through the extraction of value from the appropriated resources."[54] What they are referring to is the Marxist theory of capitalist expansion and subjugation through *primitive accumulation*. Primitive accumulation has many layers and stages in order to serve capitalism. The crux of the process entails the people's separation from their means of production and consumption into "interest bearing capital".[55] Marx has been criticized for leaving out colonialism in his analyses. Rosa Luxemburg reworked the term primitive accumulation to include a relationship to colonial policy as well as its continuous reiteration as provision for the expansion of capitalism.[56] Another naming of these processes originating from Marx, is accumulation by dispossession. A process that became a central feature of neoliberalism.[57] Introduced by Harvey in *The New Imperialism,* it described the ongoing and scalable separation of anything, human or non-human (inter)actions, still outside capitalist markets, from their originators.[58] Dispossessing the originators includes the commodification of cultural artefacts formerly belonging to the commons, slowly seceding and conforming them to market logics.[59] Importantly, regardless of how people are dispossessed, the result is that the material creators have little to no control over their materials, the conditions they are used in, or the resulting commodities.[60] Several scholars have pointed out that data relations are anything but a new configuration of capitalism, but rather a continuation of dispossession by accumulation applied to new technologies.[61] A case for this dispossession is in

[52] Couldry and Mejias, "Data Colonialism," 344.
[53] Couldry and Mejias, *The costs of connection*, 85.
[54] Couldry and Mejias, *The costs of connection*, 88.
[55] Christian Fuchs. "Universal Alienation, Formal and Real Subsumption of Society Under Capital, Ongoing Primitive Accumulation by Dispossession: Reflections on the Marx 200," *Triplec: Communication, Capitalism & Critique 16, no. 2,* May 4, 2018, 456.
[56] Rosa Luxemburg, *The Accumulation of Capital*, repred. Rare Masterpieces of Philosophy and Science. (London: Routledge and Kegan Paul), 1971, 454.
[57] Fuchs. "Universal Alienation," 560.
[58] David Harvey, *The New Imperialism*, (Oxford: Oxford University Press, 2003), 148.
[59] Fuchs. "Universal Alienation," 547.
[60] Fuchs. "Universal Alienation," 546.
[61] Couldry and Mejias, "Data Colonialism," 343.
Thatcher, "Data Colonialism through Accumulation by Dispossession," 996.

the agreement to the EULA's of platforms currently necessary for social participation, like Facebook and many of the Google services. The accumulation of large datasets stems from a performative consent request despite there being no viable alternative.[62]

In *Theft is Property,* another theorization on colonial dispossession social scientist Robert Nichols proposed a framework to understand how the native peoples of North America fought for the American land which was incrementally being subjected to the Western concept of private ownership.[63] As Nichols points out, there are two strands of colonial dispossession that are usually evoked when colonial injustices are discussed: dispossession of the self, and dispossession of land.[64] Couldry and Mejias consider the self as a territory of dispossession. I want to argue that the second type of dispossession is a more effective frame when it comes to image data. This does not mean that the other type is not in effect as well, both can happen simultaneously. Nichols uses *recursive dispossession* to describe a major historical event, the expansion of the Anglo-Saxon colony in both territory and population.[65] The native peoples had no concept of land ownership. When the Anglo-Saxons privatized the land, performing colonial expansion, they did something Nichols calls recursive dispossession.[66] It describes the act of theft and the creation of property enacted simultaneously. The act of taking the land for private use made it into property which was now stolen from the native peoples. The colonizing power generated the conditions under which dispossession is possible.[67] This also means that the people who wanted to contest this had to so in terms that retroactively validated the existence of private land ownership. The 'stolen' land infers that someone owned it before it was taken. This is the recursive aspect of dispossession. Recursive dispossession is to lose something that was not yours to begin with, because it was brought into a new system where ownership is introduced. Any issues or disputes that arise validates its establishment, can only be addressed within the newly created framework. As stated by Thatcher et al. "commodification may differ, accumulation by dispossession fundamentally entails the making private of something previously not."[68] In this lies the double-edged sword of recursive dispossession, the constitution of big data as a dispossessive source for capital encapsulated 'big data' propping up the notion of quantified data as proprietary value. Big data escapes this sense of being stolen from due to obfuscation through technical language and technological positivism. Many forms of non-privatised digital artefacts exist on the internet. Dispossession by accumulation, manages to reinvent itself continually, also

[62] Thatcher, "Data Colonialism through Accumulation by Dispossession," 996.
[63] Robert Nichols, *Theft Is Property!: Dispossession and Critical Theory* (Duke University Press), 2020. https://doi.org/10.2307/j.ctv11smqjz.
[64] Nichols, *Theft Is Property!"* 14.
[65] Nichols, *Theft Is Property!"* 9.
[66] Nichols, *Theft Is Property!"* 8.
[67] Nichols, *Theft Is Property!"* 91.
[68] Thatcher, "Data Colonialism through Accumulation by Dispossession," 996 - 997.

in the digital era through the accumulation and commodification of social relations. Still, dispossession takes particular shapes and identifying them, which this study aims to do, should aid in critically engaging with modern power relations, digital infrastructures, and efforts towards non-exploitative data relations.

§3. Methodology (1.500 words)

To determine what digital infrastructure means and can do, I take the position that media is our situation.[69] This means that the extent of agency is facilitated and constraint by the technology of media.[70] Rather than approaching this issue from a technological deterministic view, I employ technological situatedness as the guiding ontology. Technologies have great power to shape society but do not prescribe a teleological outcome. Slow-changing and ubiquitous technological infrastructures set the conditions under which individuals can operate. To lay bare the structure that resulted in the controversial outcome of generative art, and to consider its dynamics requires a method. How do the mechanics of dispossession take place? Not AI art tools nor LAION can solely be 'blamed' for engaging in data colonialism, they are an extension of the data assemblage they are embedded in and the affordances of its infrastructure.[71] In this study I will limit myself to a partial investigation as the topic is too large to discuss its assemblage exhaustively. The issues at hand are results of a practical use of the available infrastructure. In the nineties Bowker and Star developed a methodology that engages with this problem, "Some Tricks of the Trade in Analyzing Classification."[72] Investigating data relations requires the whole infrastructure to be considered, which becomes more esoteric the easier they are to use.[73] Infrastructural inversion is a methodology to reveal the processes that the infrastructures have made invisible. The article "Infrastructure studies meet platform studies in the age of Google and Facebook" by Jean-Christophe Plantin shows the use of integrated platform and infrastructure studies as a means to understand digital media.[74] The infrastructuralization of platforms has made infrastructural inversion applicable to platform research. There are notable particularities of digital infrastructures, for example its tendency to change fast. New technologies get adopted in the ecosystem and exiled or made mute over short

[69] Geoffrey Winthrop-Young and Michael Wutz, "Translator's Introduction: Friedrich Kittler and Media Discourse Analysis," *Gramophone, Film, Typewriter,* Stanford: Stanford UP, 1999: 104. Friedrich A. Kittler,
[70] William John Thomas Mitchell, and Mark B. N. Hansen, eds. *Critical terms for media studies* (University of Chicago Press, 2010), xxi.
[71] Kitchin, "The Data Revolution," 22.
[72] Bowker and Star, "Analyzing Classification."
[73] Bowker and Star, "Analyzing Classification," 33.
[74] Plantin, Lagoze, Edwards, and Sandvig, "Infrastructure Studies Meet Platform Studies," 307.

periods of time. Four themes of inquiry presented by Bowker and Star will serve as the method because it provides a guide for critical investigation of infrastructures. What follows are short descriptions of each theme and how they relate to digital infrastructure.

1. Ubiquity and interdependence: Classifications are an integral part of building an infrastructure. The classifications standardize formats and practices, making each level cross compatible and transferable, and simultaneously they depend on each other to function.[75] For example, the standardization of document formatting.
2. Materiality and texture: The aforementioned classifications are material as well. They are embedded in layered ways. Infrastructures exist of a mixture of material and culture of use.
3. Indeterminacy of the past: Narratives that seem universal are constructed, it depends on who has the power to be heard and what voices are left out. Historiography is always revising its story as voices change. For instance, conceptualizing data from a strictly technological perspective de-emphasizes the importance of its political historiography.
4. Practical politics: What appears to be universal is a result of organizational negotiations and conflicts. The design of classifications and standardizations is formed through this process which forms a practical ontology. The points of engagements that are turned into data are chosen due to their eventual possibility of capitalization.

Infrastructural inversion methods are interested in what these large-scale information systems *do in practice* rather than how they are 'really' constructed. Within this methodology LAION will be seen as a central node within the infrastructure that AI art tools rely on. The emphasis is the database because it is a point of convergence. The conditions that are set up as a result of the infrastructure takes shape in LAION. The dataset is a practical manifestation which allows datafied subjects to only be handled in a pre-set manner.[76] The texts under review are all part of the discourse network around LAION as they are in conjunction with the tools on some level. Analysing LAION as a point in a discourse network, as formulated by Kittler lays emphasis on the technological embedding of discourses.[77] In turn, this allows to consider media technology as shaping culture beyond afforded or preferred mediation but as producers of social relations. The linguistic elements of the text are part of the discourse, but Kittler points our attention to the aufschreibesystem, the notation system as radical actors that change not just the manner of cultural mediation but the cultural meaning of media itself.[78] The materiality of media technology is a "frame-constituent" and object of cultural knowledge which profoundly affects our interaction with language, ultimately producing historically different modes of literacy and

[75] Bowker and Star, "Analyzing Classification," 37.
[76] Bowker and Star, "Analyzing Classification," 36 - 38.
[77] Friedrich A. Kittler, *Discourse Networks, 1800/1900* (Stanford University Press, 1990), 259 – 260.
[78] Kittler, *Discourse Networks*, 236.

understanding. Applying a networked discourse analysis allows for a rigorous exploration on how meaning is constructed in interaction of physical, technological, discursive, and social systems.[79] Not unlike a Latourian actor network analysis, the nodes are human and non-human, and while there are differences in agency, both can influence the interaction and end-result of the network.[80] LAION facilitates observation of this manifestation through its partial open-accessibility and partial transparency in operation. Researching LAION provides an insightful entry point to analyse infrastructures in a digital ecosystem under data colonialism. Bowker and Star do not provide a distinct path to follow using this infrastructural inversion but they do provide a pattern on how to inquire using the four points.

§3.1 Corpus

Taking the four themes of Bowker and Star I have delineated the objects of this study. The goal is to provide a working framework by which to analyse the colonial logics currently at play in digital infrastructures. The tendency for infrastructure to disappear creates challenges when attempting to analyse them. Existing discussions on the topic of AI generative image models usually employ a framework that considers only a single aspect. In addition, they are either steeped in field jargon or oversimplified in their analysis. Bear in mind that I am interested in the practical manifestation of LAION as part of a digital infrastructure. The layers LAION uses and advances requires a sufficiently layered groups of sources to analyse. I will build on the approach of Mellet and Beauvisage in their infrastructural analysis of 'Cookie markets,' which captured and critically engaged with the state of internet cookies and its implications.[81] I argue that the following selection of texts and steps are necessary to map the infrastructure.

1. To investigate ubiquity and interdependence the most salient objects will be the technical structure of LAION and its supporting systems. Digital infrastructure is built on material technical systems as well as standardization protocols (e.g., URL's, and API's). By looking into what LAION does to create its final product, the personalized CLIP based image datasets, I will examine the technical documents that engage with classifications. This means publications dedicated to the development the LAION system and its co-workers Common Crawl and CLIP. LAION, as an open-source project, has published their system.[82,83] The openness of LAION's

[79] Geoffrey and Wutz, "Kittler and Media Discourse Analysis," 103.
[80] Bruno Latour, "Technology Is Society Made Durable," *In A Sociology of Monsters: Essays on Power, Technology, and Domination,* edited by John Law (London: Routledge, 1991), 108.
[81] Kevin Mellet and Thomas Beauvisage, "Cookie Monsters. Anatomy of a Digital Market Infrastructure," *Consumption, Markets and Culture* 23, no. 2 (2020): 110–129.
[82] Christoph Schuhmann, et al., "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs," *arXiv* (November 3, 2021): 1 - 5. https://doi.org/10.48550/arXiv.2111.02114.
[83] Schuhmann et al, "Laion-5B," 1-50.

materials is the reason I chose this dataset over other, private ones. While others, for example the COCO dataset, will have similar configurations, LAION comes with its own particular set of dynamics as a research project. Mapping this infrastructure using the text allows me to consider it critically. Gaining technical literacy and understanding the current set of classifications is foundational to critically engage with what is repeated by institutes, such as platforms, journalists, and developers.

2. The material textured outcome of the systems in operation is crucial to understanding the infrastructure. So far, all texts will consider the intended usage and preliminary experiments with LAION. Critically engaging with publications about LAION and the results of its downstream tasks, such as AI art, can distinguish supposed results from actual inspections. For instance, Helmond observed the platformization of the web being reflected in the Facebook's developer platform over time, as it positioned the new capabilities and purposes of the Facebook platform.[84] While this was an insightful inquiry to discover 'what platforms do,' she discussed the practices Facebook stimulated.[85]
3. The past is frequently reinterpreted, to reflect current attitudes and provide understanding of the present. Journalists are often the intermediaries between the field and the general population and have power in shaping the way people regard technologies. It shows the types of narratives and the people who get to speak on the topic. This means general reporting on these generative systems, as well as publications from journalists and bloggers that have recorded issues, possibilities and oddities that have come as a result of using AI art tools trained on LAION.
4. The fourth and final entry point, practical politics are unmistakably embedded into the texts. At this point, the analysis must move beyond description and engage with the meaning making at play in the networked discourse. Using these sources, it should be possible to map the infrastructure, critique it, and put forth an argument on what this infrastructure means for society. The analysis will not follow the four points of Bowker and Star sequentially because each point is inseparably entangled, but they are the facilitators of the findings. The analysis will follow several layers of the infrastructure. Focussing on each relevant layer by itself, which is then brought together to build a whole. The specific structure will become apparent in the next section.

[84] Helmond, "Platformization," 3.
[85] Helmond, "Platformization," 8.

§4. Analysis (6.850 words)

The analysis will consist of inspections on several layers in this infrastructure as well as a critique of emergent rhetoric around its practices. While LAION remains central, the infrastructure's systems work in tandem to power and shape LAION. It is important to understand the interplay between each system. The analysis will show that it is not one party that dispossesses, rather dispossession occurs through the infrastructure. The architecture of the internet combined with the legal position and conceptualizations of data issues found in generative image models are the result of structural processes which finds its roots in colonial mentalities and the logic of continuous accumulation. Figure 1 portrays the systems around LAION that constitute the scope of this analysis. The LAION datasets are surrounded by other actors. The systems in orange indicate software, the green denotes data, and yellow layers represent direct dependencies on tech companies. The analysis will roughly follow the infrastructure from left to right as pictured in the diagram. This is meant to mimic the chronological order of development as all nodes are dependent on their left counterpart(s) and facilitate their right counterpart(s). The breath of the diagram is based on the services LAION utilized for dataset development.[86] The nodes around generative image models are based on technical descriptions and journalistic inquiry on what the models make possible.[87]

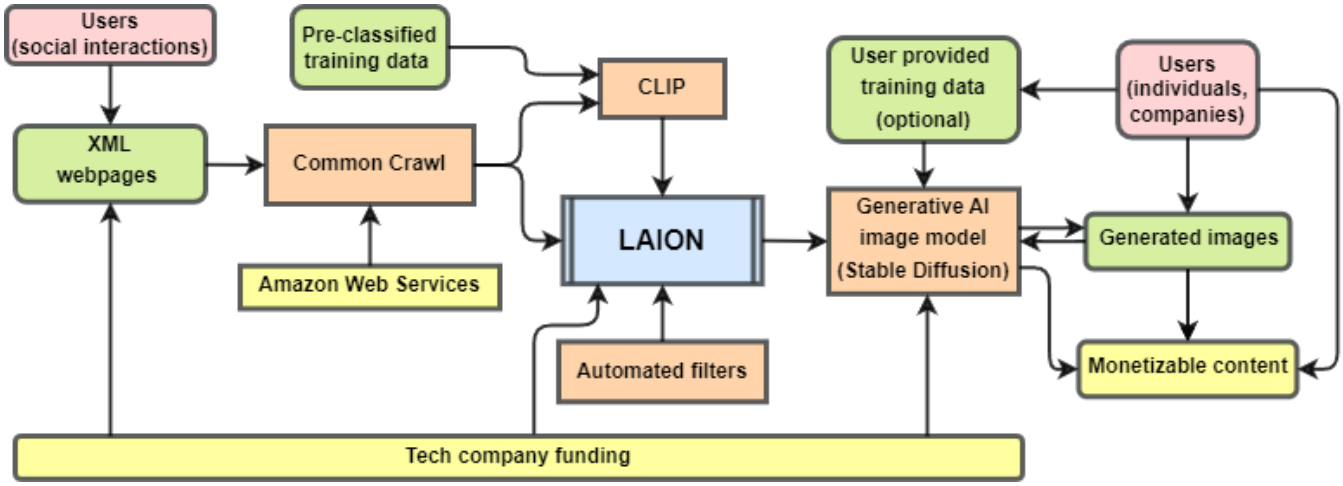


Figure 1: diagram of LAION's infrastructural position

This study is by no means exhaustive of the infrastructure. The chosen layers provide a coherent and representative perspective on data relations around LAION. The technological complexity of these systems will only be discussed in general terms. Understanding how individual systems

[86] Schuhmann et al, "Laion-5B," 13.
[87] Nicholas Carlini et al., "Extracting Training Data from Diffusion Model," *ArXiv* (January 30, 2021): 1 - 31, https://doi.org/10.48550/arXiv.2301.13188.
Andy Baio, "Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model," Waxy.org, November 1, 2022, https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/.

operate and what they do in relation to other systems, continually investigating the rationales, affordances, and power dynamics will provide insight into its practical consequences. While I will be criticizing LAION extensively in the following sections it has to be addressed first that they were chosen because it is the only openly available dataset, since it is an academic research project. Commercial actors are also using scraped data as we have seen in Meta's Make-A-Video tool which included the use of over three million YouTube videos for training.[88] Open-access, the label under which all these projects exist, also means open for tech companies to use. Moreover, these companies often co-fund this research. Despite the open and free models, they still require substantial amounts of resources to create. [89] The double-edged sword of LAION being an open-access research project is that it allows others to investigate the developments of computer vision. This could mean that LAION faces disproportionate scrutiny compared to oblique projects organized internally at commercial companies.[90] To complicate matters, these open research projects are often tied to powerful commercial entities for means of funding. This unequivocal power imbalance reveals the omnipresence of capitalism in the digital arena. Projects that are supposed to help independent developers rather provide 'free' labour and commodities to existing commercial companies since they can harness it most effectively.[91] The significance of this entanglement will be elaborated on in section §4.6. I will show how dispossession in relation to both capitalism and colonialism is a fruitful perspective to investigate data infrastructures. The infrastructure has, through its historical composition and top-down influence from platforms, created the conditions of data sharing and utilization. The development of computer vision has changed these conditions however. The conditions are now retroactively imposed on existing data, generating the friction now present, and recursively asserting it with its infrastructural implementation. The first sections (§4.1 & §4.2) engage with the curation of image and text data that make up the LAION dataset. Specifically the systems, Common Crawl and CLIP, which shows how the architecture of the infrastructure unwittingly empowers dispossession. Section §4.3 and §4.4 will investigate LAION as a digital object in depth, showing how it utilizes this affordance and how the image data are turned into something truly proprietary. The final two sections (§4.5 & §4.6) explain how the AI generative models and their images have become a point of contention as a result of recursive dispossession. I will do so by examining the discourse around its legal positioning and its efforts to justify itself as a democratizing tool.

---

[88] It used Microsoft's XPretrain dataset which is made up of clips solely from YouTube videos. Source: Hongwei Xue, et al., "Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions," *CVPR 2022* (July 2022): 1-21, https://doi.org/10.48550/arXiv.2111.10337.

[89] Daniel Schiff, Justin Biddle, Jason Borenstein, and Kelly Laas, "What's Next for AI Ethics, Policy, and Governance? A Global Overview," *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)* (February 7, 2020): 157–158. https://doi.org/10.1145/3375627.3375804.

[90] Schuhmann et al, "Laion-5B," 14.

[91] Zuboff, "Surveillance Capitalism," 78.

§4.1 Common Crawl Crawls the Commons

Common Crawl is an internet crawler, a data scraping tool designed to retrieve metatext from webpages. It operates by deploying bots, which are small programs that automatically navigate through hypertext links and gather the metadata of each page they land on.[92] Unlike with APIs, internet pages do not need to be optimized for this scraping process to occur. However, the structure of the page still influences the effectiveness of the collection of metadata. Common Crawl bots use the sitemap protocol, which is an XML standardization.[93] Similar to other systems discussed in this paper, Common Crawl is available for free to "democratise" the data.[94] They have explicitly stated that they do not limit the amount of data collected, only excluding illegal material. Some websites do not want their data scraped for a variety of reasons. For instance, to protect their content from being copied and reused or to safeguard data that can give companies a competitive edge. Privacy plays a role too, for example a crawler bot can be instructed to scrape email addresses and phone numbers. Websites can request that the Common Crawl bot not scrape their site using these embedded lines in their robots.txt file:

```
User-agent: CCBot
  Disallow: /
```

While I suspect Common Crawl does honour this request, there is no protocol that can prevent all scrapers from retrieving data from websites.[95] Still, the act of stopping crawlers requires webpage designers to locate and use specific commands (if one exists at all). This necessitates an awareness of crawlers, the Robots Exclusion Protocol, and Common Crawl as an entity. Unless these steps are performed, the crawler will copy each page's metadata. Platforms have an interest in allowing these scrapers to go over their user's pages. In its own words, Common Crawl makes 'a copy of the web,' and boasts the recording of metatext of over 50 billion web pages. Common Crawl's Terms of Use specify that they have no control over the content of the web and therefore are not responsible for the contents of their dataset.[96] Recording as much metadata from the web

[92] Alexandra Sasha Luccioni and Joseph D. Viviano, "What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus," *ArXiv* (May 21, 2021): 1, https://doi.org/10.48550/arXiv.2105.02732.
[93] "Sitemaps XML format," Sitemap protocol, Sitemaps.org, accessed March 17, 2023, https://www.sitemaps.org/protocol.html.
[94] "Can't Google or Microsoft just do that?" Frequently Asked Questions, Common Crawl, accessed March 17, 2023, https://commoncrawl.org/big-picture/frequently-asked-questions/.
[95] "What is Data Scraping?" Cloudflare, Cloudflare.com, accessed April 28, 2023, https://www.cloudflare.com/en-gb/learning/bots/what-is-data-scraping/.
[96] "3. Content disclaimers and restrictions," Terms of Use, Common Crawl, accessed March 17, 2023, https://commoncrawl.org/terms-of-use/full/.

as they are able to is not a neutral action and it enables processes that would be impossible without it. It demonstrates the naturalization of data appropriation since it treats the data as free, similar to a natural resource, as a by-product of social interactions. Doing this however, requires a specific technical, ideological, and cultural condition to be on-going.[97] It uses the technical architecture of the internet, the XML scripting, the standardization of files formatting, and its programmability, to extract data at a large scale without being noticed by the host. I do not claim that Common Crawl's developers ever intended to dispossess anyone of their data. However, it is the accumulation of large quantities of data that eventually leads to dispossession. If the data collecting was not valuable, they would not do it. As a 'copy of the web' it would entail an astronomic amount of data, which needs to be stored. Amazon hosts the dataset for free with their Open Data Sponsorship Program.[98] Amazon is offering nigh unlimited storage for datasets, which is an attractive service for these kinds of projects. Because this way they can save a considerable amount of money. Again, the investment of commercial entities underlines that these non-profit projects benefit them. It creates a mutually beneficial relationship, the implications of this will be explored further in §4.6.

From Common Crawl's dataset LAION copied image and alt-text data, first 400 million pairs and later expanding to 5 billion, called LAION-400M and LAION-5B respectively. LAION did not do anything revolutionary in a computational sense and yet it facilitates a key point in the development of AI generative image models. The effective text-image association, which is the processing that was performed on this dataset is made with CLIP, which I will discuss in the next section.[99] LAION-5B has enhanced the CLIP dataset by processing more images and incorporating multiple filters to improve its usability. The amalgamation of various systems within this infrastructure is manifested in LAION. As an intermediary, LAION functions as the final stage between data preparation and implementation, involving the selection of datasets. Nevertheless, it is imperative to comprehend the workings of CLIP, as it contributes to the issues that have arisen in the course of its utilization.

[97] Couldry and Mejias, *The costs of connection*, 89.
[98] "Common Crawl," Aws marketplace, Amazon Web Services, accessed April 19, 2023, https://aws.amazon.com/marketplace/pp/prodview-zxtb4t54iqjmy?sr=0-1&ref_=beagle&applicationId=AWSMPContessa#resources.
[99] Schuhmann et al, "Laion-5B," 1 - 2.

§4.2 CLIP: Combining Images and Text for AI

Contrastive Language Image Pre-training (CLIP) was trained using data with images coupled to text (image, text pairs).[100] The aim of the model is to accurately retrieve an applicable set of photos to a query and vice versa. For instance, when one queries 'horse' it should only provide pictures of horses, even those that did not come pre-labelled as such. Vice versa if you put in an unlabelled image of a horse the model will recognise it as a horse and retrieve images from those 'classifiers.' Classifiers are a little difficult to pin down in this case because the model is built with Natural Language supervision which means that each classification is made by a natural language model, or based on words and sentences present in the training set, but what those are and how they are related, is decided by the machine.[101] This is a step between the classical supervised AI models and the other increasingly popular unsupervised AI models. Supervised models are trained with pre-set classifiers and the machine learns what images fall into which classification. The drawback is that it requires many resources to hand-label images and it struggles with more complicated sentences.[102] Unsupervised models do not have these drawbacks and are for example instructed to create their own 'clusters' of related images based on parameters it has learned itself. The issue with this method is that it cannot name the created clusters and that it generates clusters that are undecipherable to people.[103] Natural Language supervision could be seen as a semi-supervised model where all the classifiers are created from text in the dataset by the model, allowing for a wider and more flexible set of categories. It does so by finding how visually similar images labelled 'horse' are to those not labelled 'horse.' As is characteristic of projects in the AI field, CLIP was built using a vast volume of data. But unlike LAION, it is not so clear where CLIP received its training data. They state to have used a mix of private datasets and scraped internet data, but do not specify the scrapers. They state they use "a somewhat haphazard collection of 27 datasets" as scale was their greatest concern.[104] For my purposes it is crucial to consider the origin of data. The classifiers in the dataset are also derived from scraped data, which may not just cause problems with mislabelling, but speaks to the believe in up-scaling as the provider of quality.[105] This leads to how LAION can in some respects be detrimental. The

[100] Alec Radford et al., "Learning Transferable Visual Models From Natural Language Supervision (CLIP)," *Proceedings of the 38th International Conference on Machine Learning in Proceedings of Machine Learning Research* (2021): 8748, https://proceedings.mlr.press/v139/radford21a.html.
[101] Radford et al., "CLIP," 8749 – 8750.
[102] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," *arXiv* (October 5, 2021): 10, https://doi.org/10.48550/arXiv.2110.01963.
[103] Jean Burgess et al., "Critical Simulation as Hybrid Digital Method for Exploring the Data Operations and Vernacular Cultures of Visual Social Media Platforms." *SocArXiv* (November 8, 2021): 5 – 6. doi:10.31235/osf.io/2cwsu.
[104] Radford et al., "CLIP," 8756.
[105] Luccioni and Viviano, "What's in the Box?" 5.

following section will discuss what LAION does after it combines the data provided by Common Crawler and CLIP's ability to combine images with words. As well as what principles LAION adheres to.

§4.3 LAION: Pitfalls of Multimodal Datasets

Thus far, the data is being processed but has yet to bear result or take a particular shape. Recursive dispossession makes something previously unowned into property. It is hard to pin down precisely when the image-text data turns proprietary. Not Common Crawl nor LAION claim ownership of the images in their dataset. LAION is a free and open-access image database made by a small non-profit team of researchers based in Germany. The database is open to explore using a search tool on the website. This is only to create familiarity with the database and fidget with the settings.[106] For example, you can query a word with a set image resolution to see what types of images LAION will provide. Once you have decided the parameters of your desired dataset you can request it through the website. The code of LAION is published on GitHub and anyone can sift through the 5 billion captioned images. Within this project there are structural processes that decide how the images are labelled, sorted, and categorized. For instance, LAION uses several algorithms to categorize images based on resolution, aesthetic value, and safeness (pornography and gore).[107] While minor pieces of information (e.g., log in times, Ip-tracking) are more difficult to imagine as appropriated data. Images ostensibly exist in a different category, that of human production, not computer registration. People use social media platforms to display their images to other users. Often unbeknownst to them their images are being scraped and copied for these datasets.[108] Even under EU law's 'right to be forgotten,' it is near impossible to track who has copied and has been distributing one's images. Opting out, so to speak, is practically impossible. LAION does have a channel to request removal of data. Like with Common Crawl, this puts the responsibility on the user to revoke consent after the fact.[109] A priori consent is not on the table. Apparently, the naturalization of data removed the need for a priori consent, since the data is merely a resource. This section will investigate what image data can do when it is treated as such, and how it, in its accumulative project mirroring historical colonialism dispossesses people from their own visual culture and means of representation.

---

[106] "Search demo," LAION-5B, LAION, accessed February 19, 2023, https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn.laion.ai&index=laion5B-H-14&useMclip=false.
[107] Schuhmann et al, "Laion-5B," 2.
[108] Andy Baio, "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator," Waxy.org, August 30, 2022, https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/.
[109] "7. Your Rights," Privacy Policy, LAION, accessed April 19, 2023, https://laion.ai/privacy-policy/.

LAION's dataset exists to provide vast volumes of image-text pairs for the development of AI art tools. The illuminating research by Birhane, Prabhu, and Kahembwe, "Multimodal datasets misogyny, pornography, and malignant stereotypes," goes into the systemic problems of 'multimodal datasets.' [110] They denote alarming aspects of this technology. Despite LAION-5B citing the text, it is hardly responded to.[111] Multimodal refers to the three types of data that makes up one entry: the image, its description (alt-text) and its metadata. Through their quantitative and qualitative analysis of the LAION database they discover insurmountable issues at each step. They provide ample evidence that these databases are unsafe to be employed in any area other than research. Their concern lays mainly in the consequences for downstream tasks when LAION is primed to amplify sexualized, misogynist, racist and euro-centric notions.[112] Why is this so? The answer lies in the alt-text. Alt-text is a textual element in the webpage script that is associated with an image. It was initially created to allow blind people using assistive technologies to read the contents of the image. Therefore, scholars and accessibility advocates have monitored the quality of these alt-texts.[113] However, there are significant issues with the use of alt-text as image descriptors. Firstly, only a small proportion of online images are accompanied by alt-text, and its quality is most often poor. Given the dynamics of the internet, it is crucial for websites to be easily discoverable by search engines. Higher percentages of images with alt-text on a site optimizes appearing on search queries. Quality issues include "missing important information, not being descriptive enough, resorting to stereotypical and offensive descriptors, being over descriptive (including filenames and special characters) or misrepresenting images."[114] Bias, exclusion, and systematic invisibility are frequent topics within postcolonial literature, as they are driving forces for Othering and continued marginalization. The reiteration of these issues is not only a reflection of current cultural ideas, but also an effect of colonialism. The digital exploitation reduces everything to be in service of profit generation, and the people at the margins receive the worst drawbacks, widening injustices caused by dispossession.[115]

Figure 2 provides an example of the type of images prevalent on the web. It is a screenshot of LAION's search demo which is intended to familiarize potential clients with the dataset.[116] The captions shown are not the original alt-text but the CLIP classifiers (keep in mind that these classifiers are partly based on alt-text).

[110] Birhane, Prabhu, and Kahembwe, "Multimodal datasets."
[111] Schuhmann et al, "Laion-5B,"
[112] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 14.
[113] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 2.
[114] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 2.
[115] Couldry and Mejias, *The costs of connection*, 108.
[116] Romain Beaumont, "LAION-5B: A New Era of Open Large-scale Multi-Model Datasets," LAION blog, March 31, 2022, https://laion.ai/blog/laion-5b/.

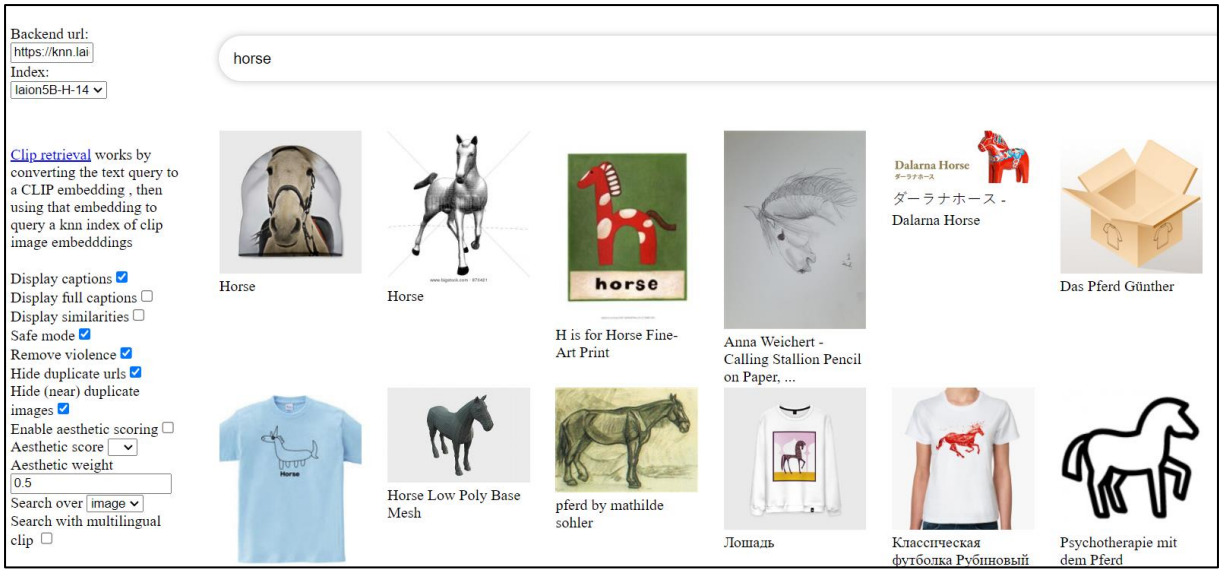


Figure 2: Cropped screenshot of LAION's search demo with the query 'horse.'[117]

Regarding the filters, one can add to cultivate their dataset, the aesthetic score filter is crucial. High quality images are more useful for training AI art tools. A source study on LAION's 12 million image dataset (called LAION-Aesthetics v2 6+) found that half of the images were sourced from 100 domains. Pinterest was the largest, followed by print selling website, Fine Art America.[118] Most subsequent websites were social media platforms, shopping sites, and stock photo libraries. The establishment of the workings of this infrastructure leads me to the conclusion that the extraction and reapplication of meaning to images is greatly depersonalized. What I mean is that the data these domains are reduced to fodder for AI training. Its content is only divided in a handful of categories like 'NSFW' or 'Aesthetically a 6'.[119] Dispossession by definition leaves the producers, in this case users, with little control over their own materials, and the shape the commodities that are derived from it are in the hands of capitalist interests. When this happens to images, it has consequences that reaches beyond the non-consensual processing of data. Additionally, it appears that in order to consistently quantify social relations one must resort to these unreliable, unrepresentative, and unintended traces of social practices that were never intended for that end. Regrettably, the same can be said for the other aspect of the matter, i.e., image data. None of the scraped websites have made a conscious decision to release their data for the purpose of training; the low quality of the alt-text is indicative of this. Although the addition of alt-text had various objectives, training a linguistic AI model was never one of them. The same applies to images, as Birhane, Prabhu, and Kahembwe have noted. A single image can contain a wealth of information, equivalent to "a thousand words," which can hardly be conveyed

[117] "Search demo," LAION-5B, LAION, accessed 05-04-2023, https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn.laion.ai&index=laion5B-H-14&useMclip=false&query=horse.
[118] Baio, "Exploring 12 Million of the 2.3 Billion Images."
[119] Schuhmann et al, "Laion-5B," 31 – 35.

semantically.[120] The loss of control over the semantic and visual meaning over the data as a direct result of this process constitutes a dispossessed quality. As I will show in the next section, with large volumes of data, economics supersede ethics. Yet LAION and by extension other AI art tools are constituted by the image-text data they extracted. LAION, however, is a university (of Munich) ordained project and therefore does justify its existence beyond profit.

LAION's database does not differentiate between sources but is by all accounts aware of the negative impact of working with scraped datasets. "We consider this dataset a research artefact and strongly advocate **academic use only** and advise careful investigation of downstream model biases (bold in original)."[121] They say the research community should work together to create open and transparent datasets and procedures for model training. [122] They state that to address this challenge, a large-scale public image-text dataset with over 5.8 billion pairs and additional annotations has been introduced. This diverse dataset can serve as a starting point to ensure balance and select safe, curated subsets for specific applications.[123] Their point is twofold. Firstly, most of the image-text pair datasets are private and thus researchers are unable to investigate its (potentially harmful) content, therefore we must have a competitive transparent dataset that can be monitored. Secondly, they pose that scaling the database up will provide the ability to curate the dataset to be more inclusive. The next section calls the claims into question. We have established the technological components of the digital infrastructure under study and LAION's position within it, thus we can move on to consider the idea of scale put forth by LAION.

§4.4 Noisy Data: Dispossession and Propertization

By own admission, scale is LAION's biggest advantage compared to other datasets.[124] This paragraph considers problematic notions around scale, and how it risks misconstruing or erasing marginalised people and inexplicit concepts. The basic idea is that with an increase in scale, accuracy will increase as well because noise is reduced. 'Noise' is the loaded term used when describing data that is wrong in some way. For instance, in CLIP many images metatext will be wrongly labelled and will steer the self-learning model in decreased efficiency.[125] Large datasets cannot fully remove the noise and it is in many ways difficult to assess the percentage of data that is noise. Regardless, there is a common notion that more data means a decrease of the impact of

[120] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 13.
[121] Schuhmann et al, "Laion-5B," 12.
[122] Schuhmann et al, "Laion-5B," 11- 12.
[123] Schuhmann et al, "Laion-5B," 3 & 12.
[124] Schuhmann et al, "Laion-5B," 1.
[125] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 6.

noise on the model.[126] This logic says that even if the percentage of 'noise' remains stable the correct results will persevere with more data. The machine will learn to ignore this bad data by itself, it will filter out the noise.[127] While for some machines this model has been successful, when it comes to CLIP'ed images the notion has been contested. Since they gathered their training data from pre-labelled sources and the shaky alt-text.[128] For instance, Common Crawl has received criticism for its propensity to contain problematic content, especially high counts of hate speech (estimated between 3.5 and 19 percent).[129] Most critics of these models are concerned with the consequences of using internet content as the linguistic and visual mediator of cultural ideas, since it is proven to contain much of humanity's more harmful expressions.[130] What is truly at the core of this issue is the confounding belief that human language and expression can be quantifiably represented and that all that is necessary is enough 'good' data.[131] Crucially, there is never enough data and more will always improve the models. Therefore, data must be infinitely collected to weed out the imperfections in the system. 'Scale' is the cure-all for the models' current shortcomings which is also the very thing justifying increased gathering. Scale improvement might work for a simple entity like horses but does not allow for the full breadth of concepts, so bias and invisibility will reoccur.

Achieving scale necessitates accumulation through dispossession. Despite the idea that images on the web are open for utilization, as many claim, the consent to use these images for LAION's end is debatable. As infrastructures tend to disappear, it is especially true for data flows, as every processing step the data undergoes will forgo informing or reimbursing the originators of the data.[132] As the internet became an integral infrastructure for societal participation, and social life users entered an (implicit and reluctant) agreement with the platforms' EULA's which give the platform certain rights over users' generated data.[133] However, the content of these agreements are nebulous for most people. But there did exist a social reality for them, one in which their images could not be used to train AI models, as the technology did not exist at the time the EULA's were agreed to. What data is and can do have changed, and therefore the terms of the agreement changed as well. However, these terms are not being renegotiated with the users. The new data extraction methods are retroactively fitted over the existing agreement. The valorisation of LAION bringing about the economic exploitation of the image data. The apparent

[126] Alex Hanna, and Tina M. Park, "Against scale: Provocations and resistances to scale thinking," *arXiv* (November 20, 2020): 1. https://doi.org/10.48550/arXiv.2010.08850.
[127] Christoph Schuhmann, et al., "Laion-400m," 1.
[128] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 7.
[129] Luccioni and Viviano, "What's in the Box?" 2.
[130] Luccioni and Viviano, "What's in the Box?" 2 – 3.
[131] Hanna, and Tina M. Park, "Against scale," 3 – 4.
[132] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 302.
[133] Fuchs. "Universal Alienation," 546.

contradiction of dispossession lies in people experiencing the Open Web as a place for social interaction. Selfies are theirs, but not one's property as anyone is technically able and free to do whatever they desire. Anyone can download and remix the images. When people upload an image to Instagram for example, there is no expectation that the unique URL leading to that page is the user's property. The scrapable data is not owned by anyone, but once it is scraped and reused it has been turned into property and therefore it is now stolen. Artists, who claim to have been stolen from, have been placing their art online for a long time. Yet 'suddenly' they consider themselves victims of theft due to these new and expanding proprietary relations.[134] Returning to data classifications, image data is not classified differently than the others, like text or timestamp data. However, extraction of image data has a closer connection to people's social relations as they see their own work reflected in the generative AI model's output. The infrastructure dispossesses the makers of the data through accumulation of it by Common Crawl, after which LAION uses it for ends that the makers were not aware of, losing control over the image. Additionally, with the constitution of the Open Web, there are no effective means of regaining control over one's data. To date, retracted datasets are in wide circulation and are regularly being referenced by machine learning research.[135] Trying to remove one's data from the world is futile. People's separation from control over their own images is obfuscated through rhetoric that normalizes and justifies the process. Two discourses I think are most important to discuss further. The first is discourse about intellectual property as it lays bare the shortcomings of the legal system and shows how those who feel stolen from recursively validate their dispossession. The second is a further consideration of the democratizing rhetoric under data colonialism.

§4.5 The Proprietary Status of Image Data

Beyond what these models are capable of, they also occupy a nebulous legal place which is currently unfolding and according to legal speculators could go either way.[136,137] Underpinning this problem is property rights and the multiple legal cases that are currently developing on this topic. For instance, Getty Images, a stock photo company, is suing Stability AI, the company behind Stable Diffusion over Copyright infringements after the generator reproduced Getty's

[134] Cloe Xiang, "Artists Are Suing Over Stable Diffusion Stealing Their Work for AI Art," Vice.com, January 17, 2023, https://www.vice.com/en/article/dy7b5y/artists-are-suing-over-stable-diffusion-stealing-their-work-for-ai-art.
[135] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 14.
[136] Dee, "Examining copyright protection of AI-generated art," 31.
[137] Jessica L. Gillotte, "Copyright Infringement in AI-Generated Artworks," *UC Davis Law Review* 53, no. 5 (June 2020): 2655-2692.

watermark.[138] Claiming that the model used Getty's images without paying for them, ergo the watermark will show. Proponents of these models put forth the fact that the images are not stored in the model. Simply put, a generative image diffusion model is built up of parameters (numerical variables) that have been constructed through processing the images, which is the training itself. AI proponents compare this to human memories of images. The same way a person does not need to have an image of a horse in eyesight in order to draw one, the AI does not have any images 'on hand' to refer to when generating an image.[139] The parameters are vast, amounting to hundreds of millions. But the model only acts once a query is performed or learns when more training material is added. How a human learns and memorizes visually is different to how a deep learning model does it as the brain never stops 'processing'.[140] I mention this because the term to describe these types of models is 'neural' networks. However, they are computational models and any comparisons to neurology is symbolic only, referring to a general sense of complexity and capability rather than being a direct comparison. Nonetheless, terms from human cognition are employed widely throughout the field of artificial intelligence, including memorization. The 'memorization' frame is used to evade responsibility of reciprocity when training these models on an artist's images. Claiming that the model did nothing more than see the images not unlike a person scrolling an artist profile and recreated the style based off the memory of it.[141] However, the capabilities of visual neural networks have altered the proprietary status of image data epistemologically, something the courts are not guaranteed to consider. Regardless of whether models directly 'contain' the training data, the image data was necessary for the creation of it since the data decided its parameters. This is where the legal ambiguity stems from. Traditionally, infringement on intellectual property required the redistribution of proprietary materials which these AI models (almost) never do. Yet it needs the material to exist at all and it perpetually continues to shape the model's output.

Individual systems absolve themselves of legal responsibility by routinely relinquishing responsibility to each other or the end user. Common Crawl understands it might copy and provide data that is under copyright, and it places the responsibility of notifying Common Crawl on others to request deletion via email or over post mail.[142] Since it does not profit directly by redistributing Copyright protected material, it is legally in the clear. LAION states that each image

[138] James Vincent, "Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content," The Verge, January 17, 2023, https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit.
[139] Carlini et al., "Extracting Training Data from Diffusion Model," 2.
[140] Carlini et al., "Extracting Training Data from Diffusion Model," 4.
[141] Baio, "Invasive Diffusion.".
[142] "3. Content disclaimers and restrictions," Terms of Use, Common Crawl, accessed March 17, 2023, https://commoncrawl.org/terms-of-use/full/.

is under its own copyright.[143] Each image is only to their image source URL and not the webpage it resides on, making it complicated to trace. People uploaded their images, for example their photography to Flickr under the Creative Commons licence, meaning that the material was free to be reused as long as credit to the originator was provided.[144] In 2016 Yahoo, Flickr's parent company, released all these images as a dataset for researchers to use. It proved to be a massive resource for the development of facial recognition technology.[145] Compiling these datasets that refer to each other or the IP holder as responsible has been referred to as "diffusion of responsibility".[146] The power of these licenses is effectively nullified, since there is a twofold problem in the proprietary status of the data used. Firstly, no party involved in the collection, sorting, provision, or consumption of this data takes responsibility for the precarious position of its content. Therefore, it is difficult to point to a single person or company that can be held liable for any infringement, as the law necessitates. Secondly, discerning whether using someone's data as training data is overstepping the license, because there is no attribution, is problematic as the dataset is public and each image links to their original URL image source.[147] LAION appropriates this URL by using its content, the image, skirting legal responsibility by exploiting the status of a hyperlink's content as open and unowned. This content is the core of their dataset, filtering them through CLIP and their quality filters, which valorises the data, and making it available as their own property, their dataset.[148]

Taking advantage of this ambiguity, scrapers, database curators, model developers, and generative art makers can freely create without restriction or expectations of credit or compensation of the IP holders. It is for this reason that we must turn to recursive dispossession as a specific mode of dispossession. It was true that the image data of a selfie contained in XML script was not seen as proprietary, so was the image data of intellectual property, because it was not possible to profit on said data without direct redistribution. LAION valorised this data as such it has become proprietary, conjoining the act of property making and theft. Leaving the originators dispossessed in a manner previously inconceivable and trapping them in these new proprietary relations. Holders of Intellectual property, like graphic artists, see their names attached to art they did not make. This occurrence is common since an effective manner of generating good images is adding "in the style of <artist>" to a prompt. Only a small collection of

[143] Schuhmann et al, "Laion-5B," 24.
[144] On Flickr images can be licensed under the Creative Commons The Attribution 2.0 which enables others to distribute, remix, modify, and create derivative works based on the original creation, purposes, provided that they give proper credit to the original creator.
[145] "Megaface." Exposing.ai, accessed April 12, 2023. https://exposing.ai/megaface/.
[146] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 12.
[147] "Search demo," LAION-5B, LAION, accessed February 19, 2023, https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn.laion.ai&index=laion5B-H-14&useMclip=false.
[148] Schuhmann et al, "Laion-5B," 30.

an artist's work needs to be introduced to the model for it to noticeably replicate a style.[149] There was an emerging call to not use AI art tools. Because it 'steals' from artists by recognizably reproducing their art style.[150] As I have demonstrated, the model does not partake in any theft. However, the infrastructure, from the moment the images were uploaded to directing the model to reproduce a style, has dispossessed the artists from their own work. Initially by taking something that was not proprietary, the image data in a webpage, and turning it proprietary as a result of the several processing steps. Reflected in the EU law "the right to be forgotten" and the case of Getty Images, is image data being able to be owned. This is giving the originators a new kind of ownership, and a right over their data. Yet as Robert Nichols points out, that right is only actualized when 'selling' property.[151] The now owned data trapped in the infrastructure is useless to the owner. Because a single IP holder cannot do anything with their image data in the way AI model developers can. The right is only actualized in the moment of dispossession, making it dispossession recursively valid.[152] The lack of consent given by these artists, or their estate, is retroactively justified by the generative models' insistence that it does not store the images despite the fact there was no course of action for the deprived artists to avoid their situation. With this in mind, the final section of the analysis will reconsider what the democratizing aim means within the infrastructural configuration and how tech companies use this to their advantage.

It is important to note that in diffusion models, it has been observed that the level of effectiveness of the technology correlates with the degree of memorization that occurs. By memorization they mean the degree by which the generated image directly matches an image in the training dataset.[153] Researchers have demonstrated that individual images can be recreated which begs the question how many of the beautifully generated items in AI art are copied fractures from artworks. Current countermeasures are insufficient to remove the problems.[154] The issue of copying is at the heart of the ongoing lawsuit between Getty Images and Stable Diffusion, as the model memorized the Getty Images watermark, proving it used or "copied" the images for training without licensing. This is another indication that memorization of fragments in these systems are an issue that is difficult to notice or test for in these models. The more "effective" models, were also far more likely to memorize, raising privacy concerns creating the possibility that this issue will increase rather than decrease with increasingly functional models.[155]

[149] Successful results still vary depending on the art style.
[150] Xiang, "Artists Are Suing Over Stable Diffusion Stealing Their Work for AI Art."
[151] Nichols, *Theft Is Property!"* 33.
[152] Nichols, *Theft Is Property!"* 34.
[153] Carlini et al., "Extracting Training Data from Diffusion Model," 15.
[154] Carlini et al., "Extracting Training Data from Diffusion Model," 14.
[155] Carlini et al., "Extracting Training Data from Diffusion Model," 15.

§4.6 Democratizing Data Colonialism

The last section will consider the effect of designating these projects as democratising under data colonialism. The analysis will conclude with the notion that, regardless of intent, working within the infrastructure, bolsters exploitative means of technological advancement. In terms of its ubiquity, it is undeniable that information systems, designed to handle a constant flow of data, are a crucial aspect of our society's functionality.[156] Without strict regulations, this leaves any quantifiable data that could potentially be traced back to an individual free to be created, processed, traded, and implemented without any restrictions, allowing for new ways of using data to be introduced into society without proper consideration of their consequences. Some argue that this is intended to promote innovation and progress, but it cannot be denied that it carries risks.[157] Self-proclaimed realists argue that using data in this manner is inevitable and that we must ensure that new technologies are accessible to everyone, not just to wealthy corporations. This is where the democratization rhetoric comes into play.[158] The asymmetric resources between non-profit researchers and the big tech companies are crucial to consider here. As stated at the beginning of the analysis, there are certainly benefits of open-access models and research. Transparency lends itself well to investigation and these projects are easier to hold accountable than internal projects at major tech companies. Simultaneously, these companies like Stability AI, Google, Meta, Amazon, and their contemporaries all benefit from the fact that all this research is being done as academic research. This is evidenced by how often they fund research projects. While the advancements they make are public, organizations with extensive funds are able to harness and circulate it widely.[159] Additionally, one could interpret the move towards scraping data as an attempt to cut costs rather than to find an accessible way to create an effective dataset. Despite the idea of open-access, the number of organisations and people that can understand and utilize it for large scale implementation is rather small. Additionally, it relieves responsibility of research methods to the researchers and their institutions. Indirectly, companies can avoid dealing with privacy and theft concerns by leaving it to academic institutions to assess the ethical implications of data usage, as they approach it from a scientific research perspective rather than a profit-oriented business practice. In the case of LAION, they mention on their website that LAION's datasets should only be used for research, it is odd then that they would partner with Stability AI.[160] While the software is free, Stability AI is a private company and more pointedly,

[156] Plantin, et al., "Infrastructure Studies Meet Platform Studies," 294-295.
[157] Carlini et al., "Extracting Training Data from Diffusion Model," 15.
[158] "Can't Google or Microsoft just do that?" Frequently Asked Questions, Common Crawl, accessed March 17, 2023, https://commoncrawl.org/big-picture/frequently-asked-questions/.
Baio, "Invasive Diffusion."
[159] Couldry and Mejias, "Data Colonialism," 346.
[160] Schuhmann et al, "Laion-5B," 12

users of the software intent to make profits with the generated art. LAION itself is co-financed by Hugging Face, an AI development company supported by venture capitalists.[161] However, this can be reframed by considering that Stable Diffusion, the AI art tool itself is developed by university researchers, funded by Stability AI, making the collaboration itself between researchers. These projects that aim to democratise AI development are dependent on the resources of companies and therefore the types of research that receive funding is in service of (projected future) profit generation rather than any benefit it might provide to society's people. While people's data is being used, the idea 'of the people for the people,' that the democratizing narrative evokes, gets complicated as it is inextricably tied up in the commercial interests of large tech companies. The tech companies' influence is present at each stage of the project. This is not a condemnation of LAION, or similar research works, but rather an observation on how it is currently unavoidable to depend on and work in service of these tech companies. As their power over digital infrastructure is consolidated it has become the obvious option to perform research at any substantial scale.

Dispossession takes place here not just in the frivolous engagement with intellectual property but with the propertization of data which was not shared to be used as training data. The situation is that these generative models require gigantic datasets in order to function at all. So, the image data of millions of people and the personal explicit labour of artists have added to a pool of valuable data without consent or reciprocity. During the constitution of LAION's dataset their images were reduced to generated labels, dispossessing the originators of the meaning. Natural Language models' biases are not only potentially harmful to people on the margins, it also effectively decides the parameters of meaning itself, leaving uncommon language out.[162] So far, all generative image model developers have struggled with harm reduction, often choosing to leave controversial topics out, which also erases the margin. Efforts to detoxify datasets have been largely unsuccessful. Other than an unfortunate side effect, it is an inevitable result of the logic of accumulation.[163] Extraction from these images goes beyond the content of the image itself. When Couldry and Mejias spoke of the threat data colonialism posed to the reality of the bare self, it referred to the idea that systems working on quantified abstractions of the world will decide not just the monetary value, but the content of people.[164] While still developing, AI generated images are on the path to be employed in many forms of representation. As users laud its ability to give shape to people's ideas even if they lack artistic ability, companies can reduce costs for their illustrations, and artists embrace the spirit of remix through AI generation. All these

[161] Schuhmann et al, "Laion-5B," 13.
[162] Birhane, Prabhu, and Kahembwe, "Multimodal datasets,"13 – 14.
[163] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 14.
[164] Couldry and Mejias, "Data Colonialism," 343.

employments could be valid and reasonable on their own terms, which is not my prerogative to decide, but it is leading to a growing ubiquity of AI generated images in everyday life. The recursive issue is brought forth by the existence of the functional generative models themselves that necessitates any protest to validate the image data as proprietary. If embedded image data is proprietary, users agreed to risk their property being utilized for this end when they uploaded it to the internet. This was never truly agreed to, but rather the result of infrastructural affordances combined with an historically constructed notion of open data.

The infrastructural processes I have untangled in this thesis are a display of the irresponsibility of reducing life to a quantified whole and the shaky foundations it is built on. LAION admits that their manner of matching sentences to images "fails to encapsulate the nuance and rich semantic and contextual meaning that the image or language might contain."[165] Yet, these models are present and empowering tech platforms with an infinite source of data. This is the crux of data relations that dispossesses people from their data. For example, initiatives like AGI (Artificial General Intelligence), a researcher's led community that promote the use of *all* available data to help models fully 'understand' the world. They teach models using the flawed reflection of the world from the internet.[166] Which, in another instance of colonial logics, negatively affects marginalised groups the most but they are not the only affected by this dispossession. According to Couldry and Mejias data colonialism leads to dispossession of human life in general, by which they mean that 'the self' is separated from their social life. Since their social life is constituted by a datafied version which due to its quantified and profit-serving character, cannot reflect the human fully.[167] Yet they must engage with themselves in their lives as though they are datafied. While datasets are an actor within the datafication of society, it is not an apparatus of surveillance itself, rather it is a result of the infrastructural constitution that facilitates a new form of dispossession. Image embeddings in XML (or HTML) scripts are being used for generative model training. Propelled by the logic of accumulation, the new application of image data has created one more market based on recursive dispossession. The processes discussed here are contemporary and unprecedented in its specificity. However, the recursive dispossession of image data is still the result of the same colonialist rationale and primitive accumulation, now perpetuated through institutionalized data colonialism.

[165] Schuhmann et al, "Laion-5B," 48.
[166] Birhane, Prabhu, and Kahembwe, "Multimodal datasets," 11-12.
[167] Couldry and Mejias, "Data Colonialism," 343.

§5. Conclusion (550 words)

In this thesis, I have argued that the infrastructural processes of generative image models are leading to the dispossession of people from their data in a way that is both unprecedented and a continuation of the logic of accumulation. People's images, that were uploaded to the web for a variety of social or financial reasons, are being used as a source for AI development. By scraping, processing, and filtering image data embedded on webpages, people's images are being appropriated for unforeseen ends. With no reasonable way to prevent or undo this, the image data has become proprietary but only in the moment the originators lost control over it. These data relations dispossess individuals from their data recursively, as acknowledgement of loss of ownership validates the newly realized proprietary status. So, how do LAION's datasets contribute to recursive dispossession through its infrastructural functionality? And how does it reflect the wider logics of data colonialism*?* While LAION is a central facilitator, it is the infrastructural constitution that engenders recursive dispossession. Additionally, the biases inherent in natural language models also contribute to the erasure of marginalized groups. Furthermore, representing life through generated images risks reducing life to quantifiable configurations. While the ubiquity of AI-generated images in everyday life may be inevitable, it is important to consider the implications of their development and use. Within data colonialism, the data as a natural resource will develop new data relations that intrude on territories of life which is in service of profit generation rather than 'progress' or democracy. As it stands, open-access models consolidate power of big technology companies more than they democratise AI development. As digital forms of capitalization are emerging and effecting everyday life in unprecedented ways, it is imperative to comprehend the processes behind them in an effective manner. Therefore, it is fruitful to see this use of data through the lens of recursive dispossession since it can analyse the mechanics of this infrastructure and its resulting pushback. The thesis exemplifies the importance of considering decolonial methods of resisting exploitative processes of developing artificial intelligence.

While LAION and similar datasets are necessary for the infrastructure for dispossession to be in effect, one must look beyond the single organization. It is part of a larger system and analysing it as such aids us in conceptualising alternatives. Analysing the infrastructure does carry its limitations. An infrastructure analysis is complex and multidisciplinary, making it challenging to cover all relevant aspects. As a result, the analysis may be limited in scope and fail to account for all relevant factors. A consequence of this is the tentative generalizability of the results. The analysis describes the infrastructure around LAION in this moment in time, which is liable to evolve. Other related infrastructures have their own conditions which are left out here. However, by giving insight into one specific infrastructure, it provides a framework to consider

and contrast others. Furthermore, future research should explore alternative data collection and curation processes that prioritize equitable and democratic ownership of data. Additionally, this analysis performed a glossary purview of a variety of actors' position on the proprietary status of image data. A deeper inquiry is needed to map how different actors within the data assemblage, such as tech companies, policymakers, and civil society organizations, understand and engage with issues of data and dispossession within data colonialism. The result of this analysis emphasizes the need for decolonial approaches to data, data ownership, and AI development.

Bibliography

Amazon Web Services. "Common Crawl." Aws marketplace. Accessed April 19, 2023. https://aws.amazon.com/marketplace/pp/prodview-zxtb4t54iqjmy?sr=0-1&ref_=beagle&applicationId=AWSMPContessa#resources.

Baio, Andy. "Exploring 12 Million of the 2.3 Billion Images Used to Train Stable Diffusion's Image Generator." Waxy.org. August 30, 2022. https://waxy.org/2022/08/exploring-12-million-of-the-images-used-to-train-stable-diffusions-image-generator/.

Baio, Andy. "Invasive Diffusion: How one unwilling illustrator found herself turned into an AI model." Waxy.org. November 1, 2022. https://waxy.org/2022/11/invasive-diffusion-how-one-unwilling-illustrator-found-herself-turned-into-an-ai-model/.

Beaumont, Romain. "LAION-5B: A New Era of Open Large-scale Multi-Model Datasets." LAION blog. March 31, 2022. https://laion.ai/blog/laion-5b/.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. "Multimodal datasets: misogyny, pornography, and malignant stereotypes." *arXiv* (October 5, 2021): 1 – 15. https://doi.org/10.48550/arXiv.2110.01963.

Bowker, Geoffrey, and Susan Leigh Star. "Some Tricks of the Trade in Analyzing Classification." In *Sorting Things Out: Classification and Its Consequences*, 33-50. Cambridge: MIT Press, 2000.

Burgess, Jean, Daniel Angus, Nicholas Carah, Mark Andrejevic, Kiah Hawker, Kelly Lewis, Abdul K. Obeid, Adam Smith, Jane Tan, Robbie Fordyce, Verity Trott and Luzhou L. "Critical Simulation as Hybrid Digital Method for Exploring the Data Operations and Vernacular Cultures of Visual Social Media Platforms." *SocArXiv* (November 8, 2021): 1 – 12. doi:10.31235/osf.io/2cwsu.

Carlini Nicholas, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, Eric Wallace "Extracting Training Data from Diffusion Model." *ArXiv* (January 30, 2021): 1 – 31. https://doi.org/10.48550/arXiv.2301.13188.

Cihon Peter, Jonas Schuett, and Seth D. Baum. "Corporate Governance of Artificial Intelligence in the Public Interest." *Information* no. 12, 275 (2021): 1 –30. https://doi.org/10.3390/info12070275.

Cloudflare. "What is Data Scraping?" Cloudflare.com. Accessed April 28, 2023. https://www.cloudflare.com/en-gb/learning/bots/what-is-data-scraping/.

Common Crawl. "Can't Google or Microsoft just do that?" Frequently Asked Questions. Accessed March 17, 2023. https://commoncrawl.org/big-picture/frequently-asked-questions/.

Common Crawl. "3. Content disclaimers and restrictions." Terms of Use." Accessed March 17, 2023. https://commoncrawl.org/terms-of-use/full/.

Couldry, Nick and Ulises A. Mejias, "Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject," *Television & New Media* 20, no. 4 (2019): 336 – 349. https://doi.org/10.1177/1527476418796632.

Couldry, Nick, and Ulises A. Mejias. *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press, 2020.

Dee, Celine Melanie A. "Examining Copyright Protection of AI-Generated Art." Delphi – Interdisciplinary Review of Emerging Technologies 1, (2018): 31-37.

Van Dijck, José. "Seeing the Forest for the Trees: Visualizing Platformization and Its Governance." *New Media & Society* 23, no. 9 (2021): 2801–2819. https://doi.org/10.1177/1461444820940293.

Dul, Camilla. "Facial Recognition Technology vs Privacy: The Case of Clearview AI." *Queen Mary Law Journal* (2022): 1-24. https://heinonline.org/HOL/P?h=hein.journals/qmlj2022&i=11.

Exposing.ai. "Megaface." Accessed April 12, 2023. https://exposing.ai/megaface/.

Fuchs, Christian. "Universal Alienation, Formal and Real Subsumption of Society Under Capital, Ongoing Primitive Accumulation by Dispossession: Reflections on the Marx 200." *Triplec: Communication, Capitalism & Critique 16, no. 2.* (May 4, 2018): 454-467. https://doi.org/10.31269/triplec.v16i2.1028.

Gillotte, Jessica L. "Copyright Infringement in AI-Generated Artworks." *UC Davis Law Review* 53, no. 5 (June 2020): 2655-2692.

Hanna, Alex, and Tina M. Park. "Against scale: Provocations and resistances to scale thinking." *arXiv* (November 20, 2020): 1 – 5. https://doi.org/10.48550/arXiv.2010.08850.

Harvey, David. *The New Imperialism*. Oxford: Oxford University Press, 2003.

Helmond, Anne. "The Platformization of the Web: Making Web Data Platform Ready." Social Media Society 1, no. 2 (2015). https://doi.org/10.1177/2056305115603080.

Kaltenbrunner, Wolfgang. "Infrastructural Inversion As a Generative Resource in Digital Scholarship." *Science As Culture* 24, no. 1 (2015): 1–23. https://doi.org/10.1080/09505431.2014.917621.

Kelly, Kevin. "Picture Limitless Creativity at Your Fingertips." Wired.com. November 17, 2022. https://www.wired.com/story/picture-limitless-creativity-ai-image-generators/.

Kitchin, Rob. "The Data Revolution: A critical analysis of big data, open data and data infrastructures." *The Data Revolution* (2021): 1-100.

Kittler, Friedrich A. *Discourse Networks, 1800/1900*. Stanford University Press, 1990.

Koch, Gertraud, and Katharina Kinder-Kurlanda. "Source Criticism of Data Platform Logics on the Internet." *Historical Social Research / Historische Sozialforschung 45*, no. 3 (2020): 270-287.

Latour, Bruno. "Technology Is Society Made Durable." *In A Sociology of Monsters: Essays on Power, Technology, and Domination,* edited by John Law, 103–132. London: Routledge, 1991.

LAION. "Search demo." LAION-5B. Accessed February 19, 2023. https://rom1504.github.io/clip-retrieval/?back=https%3A%2F%2Fknn.laion.ai&index=laion5B-H-14&useMclip=false.

LAION. "7. Your Rights." Privacy Policy. Accessed April 19, 2023. https://laion.ai/privacy-policy/.

Luccioni, Alexandra Sasha, and Joseph D. Viviano. "What's in the Box? A Preliminary Analysis of Undesirable Content in the Common Crawl Corpus. *ArXiv* (May 21, 2021): 1 – 8. https://doi.org/10.48550/arXiv.2105.02732.

Luxemburg, Rosa. *The Accumulation of Capital*. Repred. Rare Masterpieces of Philosophy and Science. London: Routledge and Kegan Paul, 1971.

Mitchell, William John Thomas, and Mark BN Hansen, eds. *Critical terms for media studies*. University of Chicago Press, 2010.

Mellet, Kevin, and Thomas Beauvisage. "Cookie Monsters. Anatomy of a Digital Market Infrastructure." *Consumption, Markets and Culture* 23, no. 2 (2020): 110–129. https://doi.org/10.1080/10253866.2019.1661246.

Nichols, Robert. *Theft Is Property!: Dispossession and Critical Theory. Duke* University Press, 2020. https://doi.org/10.2307/j.ctv11smqjz.

Plantin, Jean-Christophe, Carl Lagoze, Paul N. Edwards, and Christian Sandvig. "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook." *New Media & Society* 20, no. 1 (2018): 293–310. https://doi.org/10.1177/1461444816661553.

Plunkett, Luke. "AI Creating 'Art' Is An Ethical And Copyright Nightmare." Kotaku.com, August 2022. https://kotaku.com/ai-art-dall-e-midjourney-stable-diffusion-copyright-1849388060.

Radford, Alec, Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. "Learning Transferable Visual Models From Natural Language Supervision." *Proceedings of the 38th International Conference on Machine Learning in Proceedings of Machine Learning Research* (2021): 8748-8763. https://proceedings.mlr.press/v139/radford21a.html.

Rose, Janus. "Why Does This Horrifying Woman Keep Appearing in AI-Generated Images?" Vice.com. September 7, 2022. https://www.vice.com/en/article/g5vjw3/why-does-this-horrifying-woman-keep-appearing-in-ai-generated-images.

Schiff, Daniel, Justin Biddle, Jason Borenstein, and Kelly Laas. "What's Next for AI Ethics, Policy, and Governance? A Global Overview." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)* (February 7, 2020): 153–158. https://doi.org/10.1145/3375627.3375804.

Schuhmann Christoph, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, Jenia Jitsev. "Laion-5b: An open large-scale dataset for training next generation image-text models." *arXiv preprint arXiv:2210.08402* (October 16, 2022): 1- 50. https://doi.org/10.48550/arXiv.2210.08402.

Schuhmann, Christoph, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, Aran Komatsuzaki. "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs." *arXiv* (November 3, 2021): 1 – 5. https://doi.org/10.48550/arXiv.2111.02114.

Sitemap protocol. "Sitemaps XML format." Sitemaps.org. Accessed March 17, 2023. https://www.sitemaps.org/protocol.html.

Thatcher, Jim, David O'Sullivan, and Dillon Mahmoudi. "Data Colonialism through Accumulation by Dispossession: New Metaphors for Daily Data." *Environment and Planning D: Society and Space* 34, no. 6 (2016): 990–1006. https://doi.org/10.1177/0263775816633195.

Vincent, James. "Getty Images is suing the creators of AI art tool Stable Diffusion for scraping its content." The Verge. January 17, 2023. https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit.

Winthrop-Young, Geoffrey and Michael Wutz. "Translator's Introduction: Friedrich Kittler and Media Discourse Analysis." In *Gramophone, Film, Typewriter.* 101 – 118. Stanford: Stanford UP, 1999.

Xiang, Cloe. "AI Is Probably Using Your Images and It's Not Easy to Opt Out." Vice.com. September 26, 2022. https://www.vice.com/en/article/3ad58k/ai-is-probably-using-your-images-and-its-not-easy-to-opt-out.

Xiang, Cloe. "Artists Are Suing Over Stable Diffusion Stealing Their Work for AI Art." Vice.com. January 17, 2023. https://www.vice.com/en/article/dy7b5y/artists-are-suing-over-stable-diffusion-stealing-their-work-for-ai-art.

Xue, Hongwei & Hang, Tiankai & Zeng, Yanhong & Yuchong, Sun & Liu, Bei & Yang, Huan & Fu, Jianlong & Guo, Baining. "Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions." *CVPR 2022* (July 2022): 1-21. https://doi.org/10.48550/arXiv.2111.10337.

Zuboff, Shoshana. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30, no. 1 (2015): 75–89. https://doi.org/10.1057/jit.2015.5.