MASTERARBEIT | MASTER'S THESIS

Titel | Title

Digital Common Knowledge
Are the datasets used for narrow AI, like the Google Search Algorithm or ChatGTP, to be seen as a form of digital commons?

verfasst von | submitted by

Daniel Aaron Prem BA

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 642
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Philosophy and Economics
Betreut von | Supervisor:	Dipl.-Ing. Dr.techn. Dr. Erich Prem MBA

Table of Content

Introduction

Artificial Intelligence (AI) is the buzzword of our times. It is an opaque umbrella term that includes numerous functions and forms, ranging from algorithms that find tumorous cells on X-ray images, to programs like ChatGTP that reproduce our language. As this technology is developing rapidly, it gets harder and harder to understand the value chain behind the digital products of AI. This paper aims to shed light on one link in the production chain of the commodity Artificial Intelligence: the datasets that are used to train these programs.

Datasets have become a much-desired resource of the digital age, as they power all sorts of programs, ranging from Large Language Models (LLM) to art-generating machines. The datasets used for such programs can contain chats, e-mails, e-books, videos, blogs, images, sounds, music, relational data from a warehouse, or metadata like a specific IP address sending a document per mail to another IP address. Depending on the intended use case, the need for a particular kind of data arises to make an AI work (Jaen, 2024). The intuition at the basis of this paper is that at least some of the data used for AI is a communal product and resource that is not treated as such. A concept that seemed to fit this intuition is that of the commons, thus making the initial research question:

Is it possible to see the datasets used for AI algorithms as a form of digital commons?

Among other perspectives, the notion of commons[1] has been used as an economic concept to describe resources created, owned, or managed by some form of community. The term gains new relevance in the context of the current climate crisis, as shared resources and goods like air or soil are impacted by an industrial and digital global economy (Havukainen et al., 2022; Siegmann, 2022; Schwartz, 2020). But next to these so-called natural commons, the rise of the internet as a place to connect, share, or trade ideas and commodities introduced the term *digital commons*, marking the notion of commonly shared, owned, and/or produced digital spheres, products, or applications. Further following the intuition that datasets inhabit communal properties, digital commons are a perfect candidate for this thesis's endeavor.

[1] While often unintuitive, commons is both the singular and the plural thus "commons" will be used throughout this thesis.

To limit the scope of this paper, this thesis will focus on what is called narrow AI based on two examples. Firstly, the Google Search algorithm and, secondly, the LLM ChatGTP. Both try to understand language but can profit from different kinds of data. While ChatGTP utilizes (among other data) open-source datasets, the Google Search algorithm is more interested in the behavioralist data revealed when we interact in the digital sphere, as we will later explore in more detail. Their shared root of having to understand language makes the scope feasible for this paper while being different enough that the presented argument must apply to at least two variations, making the thesis more robust. Thus, the more detailed research question is:

Is it possible to see the datasets used for narrow AI algorithms like the Google Search Algorithm or ChatGTP as a form of digital commons?

The first section of the thesis will cover the commons in greater detail and their current research status, focusing on digital commons and data commons. Therefore, the introduction will present a concise summary regarding the state of research on commons, focusing on digital commons. The general understanding is that commons describe phenomena, commodities, or resources managed by (and typically benefiting) a collective (e.g., *About the Commons*, n.d.).

The natural commons, such as air or land, make up the initial core of the discourse surrounding commons. The discourse was then expended with abstract or so-called *knowledge commons* like traditions, knowledge, or institutions. A sub-group of knowledge commons can be found within the digital sphere and is generally referred to as digital commons. One of the most dominant discourses here is the one within the legal sphere: it explores digital commons to solve problems with privacy or ownership questions of personal data. Other authors have envisioned digital platforms as digital commons (Fuchs, 2021). The idea is that since people contribute, shape, and often even program platforms in a collective effort, one should see these entities as something communal. For example, the economist and commons researcher Marco Berlinguer argues that we need new perspectives on commons to grasp entities like Free and Open-Source Software (Berlinguer, 2020). Others are interested in specific datasets and their societal or scientific advantages. For example, medical journals discuss how data commons can be used for research (Piekos et al., 2021) or to gather data on the effects of bioactive substances (Tang et al., 2018), while others provide data commons in the form of accessible statistics (Guha et al., 2023).

When it comes to the definition of commons, some papers define the term in a general way of commons being owned or managed by some form of a collective [often with parameters specific to the content of a paper, e.g., Heath et al. (2021)], while others base their definition on parameters from the acclaimed commons scholars Elinor Ostrom e.g. (Mazé, 2023). In this sense, the literature on commons presents itself with two significant aspects: 1) a theoretical aspect regarding the nature of commons and what this entails for their optimal management and 2) an ethical aspect, as the commons mostly appear in the context of a community being denied something or a community suffering because commons were not optimally handled. Or the other way round, asking how commons could help a community. Often both aspects are entwined, as the motivation for a paper or book was a feeling of injustice in the context of a specific resource. For example, David Bollier wrote the popular science book *Think Like A Commoner: A Short Introduction to the Life of the Commons* (2014) while publishing alongside commons scholars like Elinor Ostrom in the anthology *Understanding Knowledge as a Commons—From Theory to Practice* (2019). Bollier provides a practically orientated definition of commons based on the assumption that there are no fixed commons but only the emergence of relations that make something into a commons. His definition of commons as a resource plus a community plus a set of rules is the essential definition of commons in this thesis. *Section I.I Common Knowledge* explains this in more detail.

However, this thesis needs a fitting framework that contextualizes the economic environment first. One capable of adequately conceptualizing the changing economic relations within the internet is the idea of Surveillance Capitalism presented by Shoshanna Zuboff.

Shoshanna Zuboff works at the intersection of economics, philosophy, and sociology, focusing on the impact of technology on social relations. Her book, *The Age of Surveillance Capitalism*, explores her ideas of a new accumulative logic of capitalism centered around a form of data that Zuboff calls behavioral surplus (Zuboff, 2019, 74ff). Zuboff argues that this new form of capital modifies behavior and perfects already existing mechanisms of capitalism (2019). Zuboff works by retracing the history of big data and how business plans changed throughout the 2000s. Her work combines two streams of contemporary popular analysis: the impact of data and algorithms on society and the changes within the logic of capitalism. The impact of big data is well explored both in popular works like *Weapons of Math Destruction* (2016) by Cathy O'Neil or academic papers like “Societal impacts of big data: challenges and opportunities in Europe“ by Cuquet et al. (2017) or “The social implications, risks, challenges and opportunities of big data” (Omoyiola, 2023).

The second popular discourse centers around privacy concerns and personal rights. While both topics are picked up by Zuboff, her core endeavor is to outline a new logic of capitalism and what power relations it entails. With this approach, I group her along works like *Technofeudalism* by Yanis Varoufakis (2024), *Le Capital au XXIᵉ siècle* (2013) by Thomas Piketty, or *The Shock Doctrine: The Rise of disaster capitalism* (2007) by Naomi Klein, all of which are works that try to work out either an entirely new form of capitalism (or in the case of Varoufakis the end of capitalism) or decipher a new emergent logic within its folds[2]. Next to the idea of adding new context to the prevailing system, Zuboff's work also falls into this kind of thinking by attempting to combine academic insight with non-academic writing. Aimed not exclusively at scholars, these forms of literature want to be accessible without losing their explanatory edge.

This goal of being accessible to a larger audience is visible in Zuboff's method, as she combines the methodology of a social scientist "[…] inclined toward theory, history, philosophy, and qualitative research with those of an essayist […]" (Zuboff, 2019, 21). While the core endeavor of her book is to describe this new logic of digital accumulation, the description is also normative, as Zuboff shows how this system is ethically problematic and why she believes it should be opposed. It is noticeable how important it is for Zuboff to convey her claims in a way that is understandable for everyday readers. At the same time, she presents her argument in detail over more than 500 pages while referencing 1201 sources ranging from news articles, business briefs, economists, philosophers, or other scholars.[3] While not covering the same length of time, Zuboff's methodology is akin to Foucault's in his lectures between 1977 and 1979, in which he describes the changing logic of governmentality, leading to a sharp analysis of neo-liberalism based on books, texts, and instructions of that time. Similarly, Zuboff combines countless papers, essays, and newspapers to flesh out her arguments, focusing less on other academic writings and more on contemporary publications and observational texts from the periods in question to work out the zeitgeist of an economic logic of accumulation and power (2019).

---

[2] Piketty for example does so by boiling down countless studies to the idea of $r > g^2$ while Klein argues that shocking people into a form of stasis is an intended part of modern capitalism

[3] While the number of quotations provide no direct correlation regarding the quality of the work at hand, it does give a good sense regarding the extend of research, especially when the work in question conducts a historical analysis of the argument and presents countless news articles and other sources to underline the argument.

The ensuing research regarding surveillance capitalism focuses on trying to combine Zuboff's work with other fields or concrete phenomena. The social worker and author Paul Garret, for example, published work investigating COVID-19 measures in the context of Zuboff's surveillance capitalism (Garrett, 2022), while others combined Zuboff with the works of Foucault (Joyce, 2024) or inspected her work in the context of social justice (Cinnamon, 2017). Upon release, there were notable positive reviews, some ascribing Zuboff's work an "[…] epochal feel." (Whitehead, 2019, 1) or that she is essential reading if one wants to understand the future (Whitehead, 2019). This sentiment also appears in the review by Naughton (2019) or the review by Julie E. Cohen, calling it a "[…] descriptive tour de force and a theoretical triumph" (2019, 240).

In their review of reviews, Jansen & Pooley (2021) work out the general criticism of Zuboff, namely the price of popularization in the form of non-academic writing and a lack of a clear scientific field. Morozov presents an in-depth critique of Zuboff that claims that her definition of Surveillance Capitalism is tautological, and her analysis and subsequent policy solution are neither radical enough nor thought out properly (Jansen & Pooley, 2021; Morozov, 2019). Basing his critique partly on Morozov, Haggart further argues that Zuboff's work cannot be seen as academic, as she leaves out too much of the relevant literature, instead focusing on historical documents as well as making "Exaggerated claims to Novelty" (Haggart, 2019, 231). The legal scholar Fleur Johns questions if Zuboff decoded a new form of capitalism or just worked out a new aspect (Johns, 2020). And while Jansen & Pooley write in their conclusion that Zuboff fails in trying to combine so much, they argue that it sometimes needs an analysis with this "[…] view-from-everywhere chutzpah […]" (Jansen & Pooley, 2019, 2849). They furthermore highlight that most reviews are positive and that even her critics preface their work with praise for the successful parts of her extensive work (Jansen & Pooley, 2019).

With varying degrees, the consensus seems to be that Zuboff captured something vital about modern capitalism, while the finer details are still up for debate regarding their relevance, accuracy, or true nature. The book was only released in 2019, so it is hard to evaluate its impact correctly. But its generally positive reception makes it so that even though it did not yet pass the test of time, it provides an intriguing and promising framework to better understand digital capitalism and its components. The decision to engage with Zuboff is furthermore backed by the circumstance that most of the criticism is directed against her policy proposals, her lack of engagement with other literature in the field[4] , or her writing style rather than her general framework and analysis of how capitalism unfolds itself online. Thus, I feel confident to use her work as the argumentative framework for this thesis.

The methodology of this paper will follow Zuboff's historical analysis and investigate how commons fit into the emergence of Surveillance Capitalism. Thus, the methodology is theoretical, taking Zuboff as the argumentative framework and the source for a driving logic of accumulation and the commons activist Bollier as a theoretical framework for the commons that is derived from practical experience. Where Zuboff argues for her Surveillance Capitalism as something that emerged at the beginning of the 21st Century, Bollier argues for his commons as a constant potential in the relationship between communities and the world (Zuboff 2019; Bollier 2014). In this sense, Bollier's commons are timeless; only its facets and variations change or vanish with time. On the other side, Zuboff's Surveillance Capitalism is specifically a product of our time, society, and available technology. Zuboff provides us with a logic of accumulation and power, while Bollier provides a logic of cooperation and resource management. What we need to find out is if the two logics can be united or, if not, what is preventing them from doing so. Thus, the final research question is:

> ***Can we see the datasets used for narrow AI algorithms like the Google Search Algorithm or ChatGTP as a form of digital commons under the definition of commons being the relation of a resource, a community, and a set of rules to manage them?***

The strength of this thesis lies in the narrow scope of having two clear examples, with the Google Search Algorithm directly tying into Zuboff's focus on Google as the pioneer of Surveillance Capitalism. Also, the idea of data commons is already established in the literature, meaning the thesis does not need to argue for their existence from scratch. Furthermore, this thesis promises exciting results, for even if the conclusion is that we cannot talk about digital

---

[4] Here it is noteworthy that Jansen & Pooley highlight how the reviews of Zuboff came from a heterogenic pool of fields, with academics of a specific field, claiming that Zuboff belongs to them (2019, 2843). This may hint at the reason for the common criticism of a lack of engagement with relevant literature, as different fields probably perceive different discourses as fundamental to Zuboff's endeavor.

commons in this context, the explanation of why they are not to be considered digital data commons provides relevant results and considerations.

The weakness of the paper lies in the theoretical framework; if one rejects Zuboff, the argumentation loses a lot of its footing. Furthermore, this paper has the limitation of being written by someone who has no explicit technical background. This means that computer scientists might find flaws in the assumption about AI that thus invalidate the ensuing conclusions. I attempt to solve this problem by basing the critical argument on one of the key operational methods of AI, the need for data for training, and not an intricate detail in the programming.

The main contribution of this thesis is twofold. A) the analysis of digital commons in a specific economic framework and B) the analysis of AI using the notion of commons. The combination of Zuboff's framework helps to better understand AI and the data it collects while enabling a better understanding of why some structures do not allow for commons.

The paper will be structured as follows.

The first chapter discusses the rich history of the commons, leading to a final definition of digital commons that will be used throughout the paper. This section outlines how commons is an interdisciplinary topic thoroughly discussed in fields like philosophy, economics, computer science, law, and ecology and how this makes it difficult to find a generally valid definition.

The second section introduces Shoshanna Zuboff's idea of Surveillance Capitalism, presenting the core aspects of this logic of accumulation. Zuboff's exploration of this new form of capitalism provides the guiding frame through which this thesis explores its research question. Finally, this section presents a brief introduction to artificial intelligence and how it operates. This section requires no technical understanding and only covers the mechanism on a surface level, but this is necessary to gain a rough understanding regarding the relevance of data for AI as well as understanding why Google's search algorithm is already an AI, for example.

Using Bollier's definition, the paper then analyses each of the three components: resource, community, and set of rules to determine if we can frame the datasets in question as digital commons. A conclusion collects and summarizes the thesis findings and provides an outlook toward future research. It also highlights the interdisciplinary potential of data commons for AI.

I: The Commons

This chapter will lay the groundwork for the paper by first investigating the rich history of the commons before taking a brief look at the inner workings of AI. The first section provides a deeper analysis of the commons and the definition used throughout the rest of the paper. First, a brief overview of the term's history while highlighting the most noteworthy shifts in its perception before the section takes a closer look at digital commons and what forms this idea can take. After that, the commons will be explored as the more abstract knowledge commons. Here, the current literature on digital knowledge commons, especially data commons, is reviewed before settling on a final definition of commons used throughout the paper that is less a fixed, static idea but rather a relational concept that is capable of adapting to an ever-changing world.

I.I. The Myth of a Tragedy

Sources of the term "commons" appeared in the 14th century, describing the ordinary populace, probably stemming from the post-classical Latin word *communes* (*Commons, n. Meanings, Etymology and More | Oxford English Dictionary*, n.d.). It was later used in the context of the *commonwealth*, describing a political community and the lands it tilled. The relevant economic connotation arose in 1833 when the economist William Forster Lloyd discussed the problem of farmers sharing a meadow for their barn animals and the resulting problem of allocation in his *Two Lectures on the Checks of Populations* (1833). Lloyd used the term commons in the context of a pasture that the people of a village collectively own. He argued that, following only their economic incentive, the herders would allow their sheep to graze the meadow uncontrolled, thus grazing the pasture barren, whereas the owner of an enclosed pasture would stop at a certain breaking point (Lloyd, 1833, 30ff.) This concept was picked up in 1968 by the ecologist Garret Hardin in his influential essay *The Tragedy of the Commons*, in which he presented Lloyd's argument in more detail.

While Hardin's paper was foremost an ecological one, tackling how growing populations handle a finite space of resources, the paper popularized the ideas of commons and their respective tragedy in economics. Hardin understood the term tragedy in the sense of the American philosopher Alfred North Whitehead. Hardin quotes Whitehead, who wrote, "The essence of dramatic tragedy is not unhappiness. It resides in the solemnity of the remorseless working of things."(1968). While not entwining the tragedy into any definition of commons (to which Hardin provides none explicitly), his paper conveys the notion that with commons, they inevitably end in a tragedy. This understanding could be interpreted as if the notion of a resource shared by rational agents inherently creates this undesirable outcome. Hardin's paper proved quite influential, providing a dominant view on commons for the years to come. Hardin himself later rephrased the term to "unmanaged commons" (Hardin, n.d.), as some pointed out that his tragedy is solved by introducing the idea that the two farmers are aware of the problem and start to manage the meadow accordingly. The author and cultural critic Lewis Hyde said the proper name for Hardin's tragedy should have been "The Tragedy of unmanaged, laissez-faire, common-pool resources with easy access for non-communicating, self-interested individuals." (2010, 44). But there remain instances where it is tricky to refute Harding, e.g., international fishing, as it seems impossible to impose personal property laws (Hardin's solution) and international governance is complex, while the incentive for each country to maximize their fishing output appears high (Ostrom, 2015, 13).

I.II. Common Pool Resources

A notable change in the discussion around commons occurred with the release of Elinor Ostrom's book *Governing the Commons* in 1990. Focusing on natural commons like fish, water, or land, her book provided an extensive framework on how such commons can be managed. (Ostrom, 2015). She based much of her findings on her travels, investigating countless commons in communities of varying sizes (Ostrom, 2015). Opposing Hardin's general conclusion that given a specific moral assumption, commons will always end in tragedy, Ostrom worked on showing how efficient commons are already governed in various contexts (Ostrom, 2015). She framed the problem as one of the power structures rather than an inherent one like Hardin did. As her focus was on practical commons, much of her book concerned itself with "common-pool resources" (CPR). These resources have two defining factors: a) *difficulty of exclusion* and b) *subtractability* (sometimes referred to as *rivalry*) (Ostrom, 2015, 29-33). The first means that it is hard or impossible to exclude someone from consuming the resource (breathing air, fishing in a sea), but that consumption or interaction subtracts from the available amount (polluting air, depleting the fish in a body of water). To manage such commons, Ostrom worked out what she would call basic design principles for effective and efficient commons. A clearly defined resource (precisely what part of the land is a community managing), rules, and structures molded to the local needs rather than the general policies or reliable institutions for larger scale commons, just to name a few (Ostrom, 2015). Furthermore, she noted the difference between CPR and collective goods. The letter lacks one or both factors. For example, the weather forecast is a public good, accessible for all, but my consumption does not take away from the “amount” of the weather forecast; thus, public goods do not suffer from free-rider problems, as opposed to CPR (Ostrom, 2015, 32f.). Another variation next to private goods (high subtractability, easy exclusion), are the toll or club goods, that share the easy exclusion with private goods, but have a low subtractability e.g., day-care centers (Hess & Ostrom, 2006).

I.III. Knowledge Commons

Ostrom's work, as well as several large academic conferences that started in 1985 (van Laerhoven & Ostrom, 2007), revitalized the field within the academic sphere, leading to many more explorations of the commons, especially those not anchored to natural resources. Emerging in the mid-1990s, the idea of "knowledge commons" was developed by scholars across the globe (Hess & Ostrom, 2006). These commons described abstract things like culture, traditions, or knowledge sets as commons. The idea was that knowledge itself could be a commons. These knowledge commons could be completely abstract, like local knowledge about a town, or anchored in institutions or technology like libraries or Open-Source Software.

Knowledge commons proved to be opaque in their definition, as they no longer described purely physical and palpable things like a lake or a forest. Even complex natural commons, like air, can be measured, while the knowledge of a community is hard or impossible to quantify. While the difficulty of exclusion remained, knowledge commons normally do not deplete themselves upon consumption (Hess & Ostrom, 2006). One person's knowledge of the knowledge commons does not detract from the amount of knowledge about the commons, putting many of the knowledge commons into a realm akin to public goods. Some knowledge commons even face the opposite, as increased consumption by others increased the quality of the commons. Economists refer to this as a positive network effect (Stobierski, 2020), while Hess and Ostrom note this as knowledge being cumulative (2006). The greater the number of users of a public telephone network, the more usage we get out of the same base resource (Bollier, 2006). The management of the natural commons itself can be a knowledge commons. How a specific local resource is managed is often well known to the residents, meaning we have natural commons in the form of CPR and knowledge commons in the form of the rules for managing these commons. Often, these localized knowledge commons become explicit only when they are lost. David Bollier brings the example of the people of the Indian village Erakulapally and their secret seeds as an example.

During the 1960s and 1970s, the Global West incentivized the large-scale farming of rice and wheat in India so that developing countries could support themselves better (Bollier, 2014). Due to a variety of reasons, this backfired in the long run, as the local ecosystem is unfit for monocultures of that kind. The solution lies in the reclamation of local knowledge commons in the form of a variety of seeds that are much better suited for the climate, as well as knowledge on how to plant them best. The local women managed to sustain themselves again by utilizing farming and harvesting strategies from the area's rich history. Bollier writes that neither the knowledge nor the seeds are seen as an economic input, as the woman forbids trading and selling their knowledge commons. (2014). Thus, the handling of the local flora as a CPR contains the knowledge regarding the seeds as a knowledge commons.

The idea of knowledge commons gained momentum, and questions of a general definition of commons appeared more complex. David Bollier wrote several essays and books on commons, arguing that "There is no standard formula or blueprint for creating a commons; that’s what examining any particular commons reveals." (Bollier, 2014). In his book *Think Like a Commoner: A Short Introduction to the Life of the Commons*, he argues that any instance of commons needs its distinct evaluation. Commons, natural and cultural, are not definable over aspects of the commons but rather over the relation they impose. Bollier writes:

> *"Commons certainly include physical and intangible resources of all sorts, but they are more accurately defined as paradigms that combine a distinct community with a set of social practices, values and norms that are used to manage a resource. Put another way, a commons is a resource + a community + a set of social protocols. The three are an integrated, interdependent whole.". (2014, 15)*

Returning to the original commons of the meadow and two farmers, we ought not to describe the meadow as the commons but rather as the interdependent relation between the meadow (resource), the two farmers (a community), and the way they plan to share the meadow (a set of social protocols). What Lloyd and Hardin framed as the inherent problem of the commons was just one aspect of the idea. Rephrasing their arguments in this framework, one can say that they suggested that a particular set of social protocols (let your sheep devour as much as possible to maximize your gains) was inevitable.

We find the general sentiment of Bollier in academic papers, for example, in Carballa Smichowski's *Data as a Common in the Sharing Economy: A General Policy Proposal*. Here, Carballa Smichowski defines commons as a) a shared resource, b) "[…] a bundle of rights that specifies how the resource is shared between the commoners […]", and c) a governance system of the resource (2016, 26)[5]. We can also find aspects of his definition in Ostrom & Hess when they talk about the Institutional Analysis and Development (IAD) framework (Hess & Ostrom, 2006). A diagnostic tool to inspect "[...] any broad subject where humans repeatedly interact within rules and norms that guide their choice of strategies and behaviors." (Hess & Ostrom, 2006, 41). Ostrom & Hess adapt this tool to better grasp knowledge commons and investigate their change and inner workings. This framework presents three relevant categories of Biophysical Characteristics, Attributes of the community, and Rules-in-Use as a baseline (Hess & Ostrom, 2006).

However, the broader definition of Bollier incorporates the idea of commons not being an inherent attribute to something, thus allowing for things to become commons. Framing commons not as something inherent but as something relational that must come to life reflects the documented vanishing of commons (privatizing a service, for example) as well as the rise of new ones (internet platforms organizing themselves). While Bollier's whole approach is generally broader than how "commons" is often used, it provides three distinct aspects, thus better capturing the interplay of factors that make out a commons while also highlighting the opinion that one cannot clearly define commons in general. Furthermore, I read Bollier as someone who draws from his experience as an activist dealing with commons in real life, focusing more on a definition that captures what he found in his work rather than a detailed theoretical construct. Additionally, much of the literature simply defines commons as something being owned or managed by a collective or community, underlining the intuitive understanding of commons as present throughout the literature. The reason for not deriving the definition of commons from a more academic source is that the definitions are tailored to the specific field of the papers, whereas Bollier argues for a general viewpoint of commons. His claim that commons must be inspected in their respective context is backed up by the already cited literature and the repeated approach to tailor commons to the respective goals of the different papers. The next question is, how does this translate to the question of data commons?

[5] Carballa Smichowski quotes the French economist Benjamin Coriat for his definition but refers to an essay that is only available in French. While an initial translation indicates no blatant misquoting of Coriat, Carballa Smichowski's interpretation of Coriat is quoted due to the language barrier.

I.IV Data Commons.

The concept of data commons, in a digital sense, begins with the notion of the internet as a commons. Designed to be an open space, originating from the open-access movement, the first internet platforms were open-access and could be edited and changed by anyone who wanted to do so (Bollier, 2014). Codes for projects were communal efforts rather than guarded secrets. The most famous example would be the operating system GNU/Linux, which was created by combining the free program suit GNU with the program Linux, which the 21-year-old Linus Torvalds built in a collaborative effort with hundreds of programmers (Bollier, 2014; *History of Linux*, 2023). Remnants of this time still exist in the way that e-mails are sent or how most websites still have an *https://* or *http://* in front of their domain name (Bollier, 2014). When companies first began to monetize this first draft of the internet, many programmers responded by creating variations of what we now call "Creative Commons", a legal form that enables other people to take something and build upon it with their work (*About CC Licenses*, 2019). Another aspect is the interplay of digital commons and the corporate sphere. Benjamin Birkinbine highlights how Free/Libre and Open Source Software are benefiting corporations, as bug fixes and updates by the community are used to improve commercial operations (Birkinbine, 2020). This intersection of the digital sphere and commons has also produced productive works with the idea of peer-to-peer production (P2P), a form of “[…] commons-based peer production […]” (Papadimitropoulos, 2020, 1) that tries to take advantage of the “[...] architectural design of the Internet and free/open-source software/hardware [...]. (Papadimitropoulos, 2020, 1) Bauwens et al., did something similar with their work *Peer to Peer: The Commnes Manifesto* (2019). Next to these digital commons, another idea could be found: that of data commons.

With the emergence of Big Data, more and more authors have investigated the idea of digital data commons, some as early as 2005 (Laerhoven & Ostrom, 2007). Since then, the topic has gained much traction, especially in the context of the law and the utilization of public data sets for research. In their 2023 paper *Data Commons*, Van Maanen et al. collect various papers regarding this topic, providing a brief overview of the current state of this discussion. They define data commons as:

> *"Data commons are communities that collectively and sustainably govern data and their relationships. This definition emphasizes the relationships and interdependencies between groups, the data that are in some way related to the group, and the various types of activities involved. This implies that sustainability relates not only to the data but also to the community involved in their governance." (Van Maanen et al., 2024, 2).*

Like Bollier's definition, one can see the focus on a relational aspect, namely that between a community and the data it creates and requires for specific tasks or activities. Furthermore, this definition highlights the aspect of governing this data sustainably, meaning in a way that is feasible for the community, not only for a brief period. An example here is the paper by Calzada et al., exploring people-centered smart cities that are developing so-called "data commons strategies" designed to bring more control over data to the populace in the context of the "Cities Coalition for Digital Rights." (Calzada et al., 2023, 1549). Other papers like *An Alternative to Data Ownership: Managing Access to non-personal Data through the Commons*. by Tommaso Fia, tries to solve legal questions of data ownership through concepts like the Commons (Fia, 2021). Fia also highlights the lack of a unified definition of commons, especially within economics and law as he defaults to Ostrom's guiding principles as a baseline before adopting them for his particular case (2021). Other examples of digital commons in the context of the law can be found in the paper regarding the de-anonymization concerns of private data by the legal scholar Jane Yakowitz, who wrote about data commons being an asset for societies, with benefits outweighing the risks of potential backtracking of private data via datasets (2011). This view is not limited to data commons but rather part of an ongoing endeavor to find a fitting legal framework, as the legal scholar Maria Marella outlined in her 2017 paper *The Commons as a legal concept*. Yakowitz's paper, as well as others, analyze the legal dimension displaying varying degrees of definitions for commons, with most encapsulating the general idea of data that belongs to a collective or is accessible to one (2021). Yakowitz, for example, simply defines data commons as "[…] the disparate and diffuse collections of data made broadly available to researchers with only minimal barriers to entry." (Yakowitz, 2021, 2).

The other dominant focus within the literature on data commons can be found in various papers on research data. This includes the paper mentioned before on data for smart cities or framing data sets as platforms for data sharing for interdisciplinary research, for example, in the paper *A Case for Data Commons: Toward Data Science as a Service* (Grossman et al., 2016). It is in this aspect of the discourse that we can find an altogether different definition of data commons, namely, as a platform that shares data. In their paper, Grossman et al. define data commons as "[…] cyberinfrastructure that collocates data, storage, and computing infrastructure with commonly used tools for analyzing and sharing data to create an interoperable resource for the research community." (2016,11). Jeong et al., define Public Data Commons as "[…] interoperable services for hosting data and computing infrastructure including software tools and applications for managing, analyzing and sharing data in a multidisciplinary research community while administrating data governance and security." (2022, 2). The three key aspects of Bollier's definition shine through in these approaches, albeit around the idea of a fixed digital platform rather than a relation that exists between the three components. Others try to adapt Ostrom's original principles; for example, Šestáková & Plichtová point out in their paper *Contemporary Commons: Sharing and managing common-pool resources in the 21st Century*, in which they analyze how these principles can be applied to urbanized and digital shared resources like a bike sharing initiative in Bratislava (2019). Alternatively, Madison et al. attempt to adapt Ostrom's characteristics for cultural commons in a legal framework (2017). Another aspect is the interplay of digital commons and the corporate sphere. Benjamin Birkinbine highlights how Free/Libre and Open-Source Software is benefiting corporations, as bug fixes and updates by the community are used to improve commercial operations (2020)

There is also a productive discourse surrounding the general purpose of data commons and how they are perceived. A paper from 2023, *The future of data ownership: An uncommon research agenda.*, by Jacqueline Hicks, argues that she wants research that "among other things, "[…] prioritizes a focus on social and communicative practices involved in a data commons rather than seeing it as a solution to a specific problem […]" (Hicks, 2023, 544). Hicks concludes that the research on data commons ought to investigate the various examples of communities that already manage their data and, by that, work on specific cases while including more ideas from non-western perspectives as well as more theories from humanities rather than trying to understand and solve the task by legislation alone (Hicks, 2023). Other authors argue that legislation alone can hardly be the answer, as the data governance might simply be too complex for a "one-size-fits-all" solution, as the economist Carballa Smichowski argues (2019).

I.V Defining the Data Commons

We have seen that the notion of commons moved from describing a natural resource to describing entire fields of science. Moving from CPR towards knowledge commons and now to data commons, the idea of communally governed resources has taken every imaginable shape. While the literature is rich, there is no unifying definition of commons. In this context, the definition given by Bollier appears to be useful and is present in different versions across the literature. ***In the following, a 'commons' refers to the interdependent ternary relation of a resource, a community, and a set of social protocols.***

Looking at the idea of data commons, one can see that the concept has been analyzed from various angles, particularly from the perspective of legal issues and data management. We saw that some authors did not define this kind of knowledge commons besides the intuitive understanding based on Ostrom's principles for natural commons. Hess and Ostrom see data as knowledge commons, quoting Machlup: “Knowledge derives from information as information derives from data.” (Hess & Ostrom, 2006, 7), thus going even broader in their definition, encapsulating every form of knowledge as a knowledge commons. Others used definitions akin to Bollier's relational structure. The idea of data as commons has found its place in EU legislation papers and various research papers (Calzada & Almirall, 2020; Zygmuntowski et al., 2021). Therefore, one might argue that the question of data commons in the context of AI has already been answered. Nevertheless, with Bollier, we can see that this is not the case, as his definition makes it clear that nothing is inherently a commons. The relation of this intangible tiple needs to be inspected to see if the data in question can already be considered commons or if it could become a commons. As the data of AI is scattered and seemingly unconnected, there is a need for an additional framework to see the connecting lines that bind it into one comprehensible phenomenon. Finally, the notion of data commons needs to be clarified, as there can also be analog data commons (e.g., analog datasets or libraries). After that, this thesis analyses what the potential resource is, if there is a community, and if the community has a set of social practices in place to manage the resource.

II New Age, New Data

The next step is to delve into the process of how the needed data is generated. This exploration of the process is executed in the philosophical-economic framework of Shoshanna Zuboff and her extensive work *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (2019). Her book begins with a historical analysis of tech companies in the early 21st century and how they learned to make data a business model. After working out this logic of digital capitalism, she elaborates on how the collected data modifies behavior and thus exerts power over people. This thesis will focus on the general logic and method of Surveillance Capitalism while holding in mind that for Zuboff this has an inherent moral and political component.

It is a common misconception that using a platform like YouTube or Facebook is free. We log on to consume content, news, and entertainment while receiving the occasional commercial. In this context, we perceive the consumption of the commercial as the "payment", meaning we watch it, and in return, we get the rest for free. In this framework, one finds the well-known sentiment: *If it is free, you are the product*. The key argument from Zuboff's work is that this view is missing the actual dynamic of the production chain at hand, leading to a misunderstanding of our role. The following section explains how Zuboff defines what she calls behavioral excess and how it plays a vital role in her view of modern capitalism. The thesis will not argue for, defend, or trace Zuboff's argumentation in detail but instead focuses on bringing across the gist of Surveillance Capitalism as the underlying framework of the general argumentation.

II.I The Discovery of Behavioral Excess

Surveillance Capitalism expresses itself by monitoring behavior, generating data that describes said behavior, and using this to optimize its own tools reflexively (Zuboff, 2019). This process is not only surveilling its participants, but also altering and changing their behavior (Zuboff, 2019.). For Zuboff, the key to understanding this process is the discovery of behavioral excess, the unexpected digital gold first found and utilized by Google[6].

[6] Officially renamed to “Alphabet” in 2015

To start, we must ask ourselves: what happens when we google something? We reveal the topic, word, term, or question that interests us. Rather per accident, however, Google realized that there is a trove of "[…] collateral data such as the number and pattern of search, phrasing, spelling, punctuation, dwell times, click patterns and location." (Zuboff 2019, 67). When search engines like Yahoo and Google first stumbled across this additional data, its economic potential was not realized. It was Google who first worked out that this excess data could be used to increase the quality of their Search engine. By using this collateral data to analyze what someone wants, needs, or means, the data increases the user experience. Zuboff describes this as a phase in which all this extra value benefited the users without any cost and was in line with Google's declared goal of organizing the world's information, making it universally accessible and useful (Zuboff, 2019, 70). But the increased quality that other engines lacked, came with no inherent way of returning the money to investors. While their promising technology was tempting enough for Silicon Valley venture capitalists to sustain Google through years of low or no profits, the dot.com bubble in 2000 required that firms like Google showed more constant and increasing profits (Zuboff, 2019, 74).

This shift led to a fundamental change in tactics for Google as they started to use this collateral data for profits. Zuboff highlights, as she calls it, an 'emblematic patent' issued by computer scientists at Google in 2003 named "Generating User Information for Use in Targeted Advertising." (2019, 77). Generating User information, or UPI (user profile information), was the first step towards capitalizing on the access data created by people using the Search engine by creating profiles of people that could then classify and predict said user. This process enabled Google to tempt advertisers with a method to allocate their marketing budget in a way that regular means of commercials never could[7]. Instead of collecting whatever excess data their engines found, Google began to hunt and explicitly create this data, resulting in what Zuboff calls Behavioral Surplus (2019, 74-82)

This Behavioral Surplus feeds algorithms tuned to predict people (Zuboff, 2019, 82f, 197-202). From anticipating what a person in a specific time and place wants to know when they google a phrase to recommending a fitting YouTube video in your feed, the algorithms use

[7] So far, we have taken Google as the main example for this new facete of capitalism but that does not mean that it only applies to this company. Rather, Zuboff outlines the history of her idea with the first appearances of its working and logic, all of which she pinpoints within the early days of Google.

the behavioral surplus to predict people (Zuboff, 2019). The predictive element no turns twofold: 1) it needs to predict users in order to better sell to advertisers, but b) it needs to predict users to gain more data from the interactions (Zuboff, 2019). It is here that Surveillance Capitalism began to modify behavior as Zuboff retraces how sites like Facebook are engineered to create addictive behavior that keeps a person on their page (Zuboff, 2019, 298-308). For example, former Facebook product manager Antonio Garcia-Martinez accused the company of predicting when users feel stressed or otherwise compromised and using this information to impact their behavior (Garcia-Martinez, 2017). While doing so, any new interaction feeds the algorithm with the new behavioral surplus, making the future prediction machine even better—a self-improving circle of accumulating data while impacting how we behave (Zuboff, 2019 93-97). Zuboff explores how, with smart-homes or autonomous vacuuming robots, the amount and kind of data is ever-increasing, allowing for newer ways of using data and exerting power (2019, 267ff.). An extreme form of Surveillance Capitalism can be seen in China's Social Credit System, in which extensive surveillance is used to punish or reward certain behaviors while gathering more and more data to find more streams of control (Creemers, 2018; Zuboff, 2019, 388-394).

The reason why Surveillance Capitalism works so well is because its tools are genuinely helpful for its users. Google is of great assistance at filtering the vast sea of human knowledge, and there is nothing inherently wrong with a social media For-You page tailored to our wants or being able to check if the oven is off. However, when a Roomba is filming a person on the toilet because it gathers data to train other AIs, this technology turns sour (Guo, 2022). Zuboff extends deeply on the dangers of surveillance capitalism, both for individuals and collectives or how social media and data can influence democratic election processes (Zuboff, 2019, 277-279). But for this thesis, the ethical dimension of Surveillance Capitalism is not of concern but rather the notion that within Surveillance Capitalism, a) data is acquired in a reflexive way by observing the users, and b) this data can be used to exert power in various forms. Thus, finding out how data is created, processed, and used is of general interest to better understand a relation that is potentially harmful (or, according to Zuboff, highly dangerous to us already).

To summarize this concept, Zuboff splits the mechanism of Surveillance Capitalism into four parts: 1) the logic, 2) the means of production, 3) the products, and 4) the marketplace. (2019, 93-97).

1) The logic of Surveillance Capitalism is to turn the resource of behavioral surplus into money by feeding it into powerful prediction machines, aka algorithms. In this logic, we are no longer the subjects of value realization or the selling product. Instead, we are turned into objects from which raw materials are extracted. The logic of surveillance capitalism turns us into the means to others' end. (Zuboff, 2019, 93f.)
2) The means of production are the algorithms that collect the vital behavioral surplus and implement it through ad placing and behavioral modification. These means of production are self-enforcing as the users continuously train the algorithm. These means of production are summarized as "machine intelligence" by Zuboff. (Zuboff, 2019, 95f.).
3) The products are the resulting prediction machines that function due to the enormous amount of behavioral excess data extracted from internet users. The capability to predict and assess makes an algorithm the product for companies, political parties, or organizations to advertise themselves. (Zuboff, 2019, 95).
4) The marketplace of this form of capitalism is what Zuboff calls behavioral futures markets. While the dominant players are advertisers, the market is open to anyone with enough purchasing power. Zuboff draws the analogy that Surveillance Capitalism is as much about advertisements as Ford's new mass production system was incidentally about cars (Zuboff, 2019. 96).

II.II Machine intelligence

In part 2) of her definition, Zuboff uses the term "machine intelligence". With this, Zuboff attempts to deal with how wobbly the idea of AI can be when she writes, "The same term may mean one thing today and something very different in one year or five years. For example, Google has been described as developing and deploying artificial intelligence since at least 2003, but the term itself is a moving target, […]" (2019, 95). One could frame Zuboff's "machine intelligence" as an umbrella term describing various forms of technology that try to mimic human intelligence. Five years after its publication, Zuboff's hunch proved to be promising as the term gets used in countless contexts, often without being defined. More of a buzzword than a technical description, the definition of AI is hard to grasp. A common definition is that of a program aiming to "[…] imitate intelligent human behavior" (*Artificial Intelligence Definition & Meaning - Merriam-Webster*, n.d.), technology that "[…] enables computers and machines to simulate human learning, comprehension, problem-solving, decision making, creativity and autonomy." (*What Is Artificial Intelligence (AI)?*, 2024) whereas Sebastian Rosengrün prefers to talk about AI more in the context of a specific area of software development rather than AI in general (Rosengrün, 2021, 16f.). In their introductory book *Künstliche Intelligenz verstehen: Eine spielerische Einführung*, Noack and Sanner further point out that there is a flowing transition between AI and computational work like data mining (2023, 26)

In this context, the main difference between AI and "regular" programs appears to be a change in instructions as well as a change in what programmers want from their work. Normally, a program should deliver precise results that remain the same, e.g., asking a calculator what 2+ 2 is should always result in 4. A video game should always have the same response to a button input and a management tool should always do the same thing given the same instructions. AI, on the other hand, works with approximations. There is no one correct answer and, especially not one correct path to an answer. Asking ChatGTP a question can result in countless answers that vary in length, content, and tone. For image-generating AI there is not a single correct way to display an apple and so on. This means one could say that AI tackles problems that have a more shaded solution than one that is solvable by tasking a program with "If I do this, you do this." To better understand this, we look at a famous example of the MNIST Dataset for handwritten numbers.

II.II.I What Number am I showing you?

Image 1. From *3.3. The MNIST Dataset — Conx 3.7.9 Documentation* (n.d.)

Most people who learned how to write and/or read can decipher all of the above images as the respective number. They may differ in shape, exact form, and style, but one can recognize them as a “2”, “3”,”6” and so on. When we talk about "learning" in the context of AI, a "2" for example, has no inherent meaning to a program. Given instructions, it interprets the symbol of 2 in a way that we want while not being able to "understand" the meaning of two. Humans learn that 2, II, and "two" all represent the same things; however, none of those three pictures have anything inherently in common for an AI. Programmers, so humans, have this understanding and do their best to translate it into a machine language.

So, how does the AI learn to recognize the two? The short answer: data and complex algebra. In his short video series "Neural Networks", the mathematician and YouTuber Grant Sanderson presents a comprehensible exposition of how machine learning works (2017). His example is that of an AI learning to read the so-called MNIST dataset, a collection of handwritten numbers always displayed on the same-sized pixel grid. It is used to build an AI that is capable of deciphering handwritten numbers. In Sanderson's example, the machine must not only be able to identify what a two is but must be able to differentiate it from a four or a zero, and so on. A possible solution is to break down the digits into compartmentalized structures that an AI can single out. For example, a 0 is always some form of an oval, round, or oblong-shaped thing, while a 1 is usually a straight line, often with another line on top of it. So, while it is hard to convey the meaning and structure of a "two" a straight line seems to be more feasible. However, the problem remains: what is a line or circle to an entity operating on the core idea of 0 and 1, On and Off. There is no entity "seeing" and interpreting a line as such. What Sanderson shows in his video is one possible solution. The images that the AI detects all have the same grid of pixels. So, one might define that specific pixels are more active than others depending on if they depict a part of the number in a specific area (Sanderson, 2017, 08:35-11:33). If certain areas light up, it could indicate that a certain sub-structure of the number is active. If several of those substructures are active, they could indicate a certain number. For example, one could dissect a “1” into one long line and a small one on top, and a six into a circle and a curved line on top. The separated structures could enable the machine to read them better. But the portioning of the numbers goes deeper than a circle and a curved line. Both nine and eight have circles with curved lines attached to them, and a circle is nothing else than a curved line bent into itself. A curved line is nothing, but an array of straight, short pixel lines arranged offset to each other and so forth (Sanderson, 2017, 08:35-11:33).

The solution to this problem is a complex mathematical approximation where several statistical and algebraic tools are combined to filter out the connecting dots. This process contains countless iterations of trial and error as the AI is attempting to solve the problem, gets feedback on its proposed solution and then adjusts certain elements in its calculation. In the next run, the AI can then see if a certain adjustment led to a better outcome or a worse one and repeats the process (Rosengrün, 2021). A big struggle behind AI is, thus, writing complex algorithms that are capable of realizing and adjusting those internal factors to achieve the desired result so that the machine can go through countless repetitions without the need for a human operator to give feedback on every single result (Rosengrün, 2021.). However, to do so sufficiently, AI needs training data. In this example, the MNIST dataset contains 60.000 images of 28x28 pixel squares that contain human-written digits from 0 through to 9[8] (LeCun et al., n.d.). While most AI must deal with more complex problems, the fundamental principle remains. Through countless iterations, the program trains by skimming datasets to find the connecting lines. The author Meredith Broussard argues that the Go-AI AlphaBlue that beat the Korean champion was not that special when we think about it, framing the victory as that of a machine that used brute calculating force to work through nearly every recorded move possible, find the pattern, and thus predict game-winning moves against a person who could spend their entire lifetime playing Go without seeing as many games as the AI did (Broussard, 2018, 36f.)

[8] All 60.000 images are labeled, which tremendously helps to train the program, as there is a built-in method to check results during the learning phase When the program trains, it will come to result that then needs to be checked. If the dataset is labeled this is much easier, as the program can just check if the predicted number is the displayed one.

II.II.II Narrow, General, and Super AI

This section aims to broaden the conception of AI and show how algorithms like the Google Search algorithm already fall under definitions of AI. As of now, the distinction between narrow AI and general AI is common in the literature (Rosengrün, 2021,19f.). More detailed distinctions concern themselves with the way an AI learns or creates its output, but these distinctions do not play a role in this thesis.

Narrow AI focuses on specific tasks, attempting to model human thinking (Rosengrün, 2021). For example, DeepBlue is a narrow AI solely capable of playing chess. Maybe the most famous and widely used narrow AI is the Google search engine, which focuses solely on interpreting a word or sentence and trying to predict what the person is looking for (Parthasarathi, 2024). General AI, on the other hand, would be a program that can apply its algorithm to various problems and fields, truly acting with intelligence but the general argument here is that this has not happened yet (Broussard, 2018; Rosengrün, 2019). Intuitively, some might think that something like ChatGTP is general AI, but it is considered weak/narrow AI, focused solely on the task of language processing. A superintelligent AI would be a system that surpasses human intelligence or at least truly mimics it, meaning it can fulfill different tasks and implement transfer knowledge. Currently, the broad consensus is that all known AI models are narrow AI focused on excelling on specific tasks (Rosengrün, 2021).

AI needs to find the connecting patterns amidst an ocean of data to determine what we want from the program. Thus, narrow AI needs data to fuel its learning, making data one of the critical resources in this field. Sometimes, we do not even need a specific goal to utilize the AI and a pool of data. Unsupervised learning is the process of tasking an AI to find patterns in each data set. The AI has no specific goal except the search for connecting lines. Rosengrün further highlights that one ought not to confuse AI with algorithms, arguing that every AI consists of algorithms or a chain of commands, but not every algorithm is an AI (Rosengrün, 2021). Finally, it is important to note that when we talk about the Google Search Algorithm or ChatGTP as AI, it is not one algorithm or "one AI" but rather a conglomerate of algorithms and AIs that form one product. Additionally, the term data now refers to data relevant in the context of Zuboff (behavioral surplus) and/or open-source data used to train an AI like ChatGTP.

III Digital Common Knowledge

As digital commons are often framed as knowledge commons, a framework that would intuitively offer itself to inspect this form of data would be IDA framing presented by Hess and Ostrom. The problem hereby is that the IDA framing assumes knowledge commons and presents a tool to better grasp them. It is not designed to determine if something is a knowledge commons. While the baseline of the IDA has a lot in common with Bollier's definition, it already assumes an “action arena” and patterns of interactions to investigate the institution at hand[9] (Hess & Ostrom, 2006, 64). Institutions in this context are “[…] formal and informal rules that are understood and used by a community.” (Hess & Ostrom, 2006, 42). This means a community, for example, is already presumed, while this thesis wants to investigate if there are any grounds for it in the first place in the context of the chosen form of data. This section will continue by investigating the three aspects of Bollier’s definition, starting with data as a resource, followed by the points of community and a set of rules.

III.I Data as a resource

Data has become a resource that is sold, traded, and used. Be it called AI, machine intelligence or just "powerful algorithms", modern, digital capitalism is thriving on this new resource. Its cost of duplication is verging on zero as it takes only the tiniest amount of energy to copy large files of data, thus making most data non-rivalrous since the copying process allows for unlimited people accessing a dataset for example. Nevertheless, it is noteworthy that both the non-rivalry and the low cost of duplication are only applicable in the context of a giant net of servers that are non-stop fueled by energy as well as hardware components, that consume energy during their production, and every time someone interacts with a data point[10]

Looking at data as a part of the value chain, we can see how data has moved from primarily being a product towards being another piece in the line of production long before the internet. In the Cambridge Dictionary, the word “data” is either synonymous with information or described as digital information used by a computer (*DATA | Bedeutung Im Cambridge*

---

[10] While the cost of coping a file on a computer appears to be 0, the act of switching on your computer, displaying the screen, ordering your computer to copy the file and so on, consume energy, thus making the act of duplication not truly zero.

Englisch Wörterbuch n.D.). Data as information can be the product of studies, observations, or thoughts and is often presented as the result. One could say that when this data is used in another study the data is used as a resource. Before the internet, these datasets were often semi-rivalrous (e.g., a limited print run of a study or book meant only a few people could have access at the same time, or one had to travel great lengths to gain access), and the cost of duplication was considerably higher (e.g., books had to be copied by hand and so on)[11]. Data or information is a resource to build and develop new technologies, models and theories, making data a valuable resource in nearly all scientific endeavors. In the realms of politics and economics, data is a much-needed resource to make optimal decisions. Like other resources, data needs to be either harvested or created, processed, and evaluated for quality before being fit for usage.

Thus, data as a resource is an accepted framing, with entities like the World Economic Forum publishing articles arguing that data should be considered a natural resource (Neri, 2020). Similarly, the IT giant IBM called data a natural resource in a post from 2019 (Smith, 2019), as well as NVIDIA CEO Jensen Huang (*Bloomberg Technology*, 2024). This perception is interesting in the context of the commons as it were natural resources that made out the first commons and are still its most prominent cases. Natural commons all share that there is no initial ownership over them. The possession evolves through the iteration of social change. But with data describing personal behavior there is much more levy to argue for an inert right over data. For this case, it is sufficient to remember that a) data is a resource, b) the relevance of this resource is constantly rising within Surveillance Capitalism, and c) some go as far as to call it a natural resource. Especially the latter bolsters the underlying assumption of data having the potential to be commons, as, speaking with Bollier, nearly every natural resource can be a commons (Bollier, 2014).

[11] Combined with a wide array of social barriers regarding who could study, read or access a specific set of information.

III.II A COMMUNITY FOR DATA

Before inspecting the idea of the users' creating data as a community, one must ask what a community is. Commonly, a community is described as a collective of people connected by a shared sense of belonging or togetherness. This could arise from a local factor like being born in the same village, a shared hobby, or attributes that heavily define our experiences, like skin color[12] (*Community Definition & Meaning - Merriam-Webster*, n.d.). Looking back at the knowledge commons, one can see how some commons, like culture, already count thousands of people scattered across the globe. The community for a hobby, connected via the internet would be another example. As with all groups, the larger the group, the more complex its management becomes. In the context of commons, the size of the commons and the size of the community play a vital role.

For example, air is one of the most extensive natural commons that postulates one global community. The problem is that the larger a community, the higher the chances it consists of several sub-communities. In this case, countries, for example. While we, as inhabitants of the earth, would be the larger community, it is the self-perception of a national or at least a cultural community that predominates. Geographical factors also shape a sense of community, like living in a city or inhabiting a particular type of land (valley, mountain, lake, and so forth.) In this context, one could already consider that there is no global community but countless national or cultural ones that share the need for breathable air. The next problem is one of responsibility. In a disjointed global economy, it is easy to demand fresh air while ignoring and accepting one's own role. We want fresh air but also a strong economy that constantly produces CO^2 emissions. Ostrom noted that commons profit from a clearly defined group of people that feel connected to it and thus manage the commons (Ostrom, 2015, 188). While she highlighted smaller communities as generally faring better in managing a CPR, she noted clearly defined boundaries as the most relevant factor (Ostrom, 2015, 90ff.). Additionally, a smaller community makes control and potential punishment easier. If a small village is managing a lake, for example, most people know each other, thus, the people who monitor the commons are often not extern to the village community, and breaking the rules is thus riskier (Ostrom, 2015, 59ff.; 186f.). But the need for punishment normally arises in the context of CPR as these commons

[12] Something we can see when we talk about the African-American community for example.

are rivalrous, and a repeated rules breach would drain the resource, whereas digital commons do not exhaust upon consumption. This means the idea of the community punishing or managing the resource is more applicable in the sense of containing the data and not selling or giving it away freely, like the Indian woman keeping their seed knowledge. Furthermore, smaller communities make it easier to create a sense of belonging and purpose. We can see this in cities, for example. While people might perceive themselves as part of a city, they are more likely to see themselves as part of a neighborhood, thus potentially putting the needs of the neighborhood above the needs of the city. Looking at the data commons in the context of Zuboff and AI, the same problem arises: the users who generate and archive the data are scattered across the globe, making it hard to see any form of "community". This effect is amplified as the data extraction imperative is omnipresent throughout the internet. The internet has shrouded the globe in an invisible net that summarizes us as a global community. But with data in the context of AI, the problem is that they are not perceived as potential commons. While Ostrom's points are directed at CPR and not at digital commons, the aspect of having a clearly defined boundary applies nonetheless. Ostrom argues that both the actors who are allowed to draw from the resource, as well as the resource itself, should be clearly defined (Ostrom, 2015, 91f.). This aspect is clearly lacking with the selected datasets. Neither do we have a clear boundary regarding who should be part of a community, nor do we have a clear boundary regarding what data we are talking about. The latter point is amplified by the opaqueness of the companies and states operating online, as well as the missing legal leverage to gain access to all accumulated data. Furthermore,

Going back to the air example, one can see how air postulates complex socio-political problems and why we ought to care about it. However, with data, people are scattered and thus have no intuitive reason to feel connected in this particular sense. When it comes to larger commons, Ostrom further postulates that one needs "nested enterprises", meaning a combination of smaller communities and larger, established institutions and governments to properly manage the commons (Ostrom, 2015, 101f.). The digital commons harvested by Surveillance Capitalism most likely fall under this category. And while there are larger institutions and governments in place, there is nothing that connects them to smaller communities. In her work about knowledge commons, Ostrom points out how a homogenous goal can create a community without people ever getting to know each other, giving the example of the open-source movement (Hess & Ostrom, 2006, 49). Here, a big problem with behavioral surplus data as commons arises, as there is no goal or purpose that could unite people, whereas the Open-Source movement has clear goals and tasks,

Throughout her book, Zuboff highlights how companies like Google ensured that their means of production remained unknown to the public (Zuboff, 2019). Bollier, on the other side highlights how commons are not inherently commons but become commons in a process of awareness and action by the community (2014, 15). This means that if a set of people is not aware of the resources they produce, they are unlikely to attempt to manage them, and unlike a large sea near a village, data remains unnoticed. Looking at open-source data like Wikipedia, the communities that managed them could not anticipate that this free data would eventually be used to feed new AI algorithms with questionable impact on society. Combined with the current opaqueness of how AI operates, people cannot really assess what is happening to the resources they create and who profits from it, especially as the non-rivalrous nature of digital data makes it so that there is no point of realization that someone meaningfully harvests your data. It is here that we see the strength of Bollier's definition. Wikipedia is a knowledge commons in the context of people managing it to share free information for everyone. But in the context of this data being the resource for a private company like OpenAI, it fails to be a commons as it is not managed in this particular context. Due to the non-rivalry of data, it can be both commons and not commons, depending on the relation at hand.

A potential answer to this problem could lie in how data is processed. While the singular users providing the data are unaware of their role and their fellow "community", the corporate algorithms have categorized the users into "communities". Be that location, culture, hobby or other markers, the whole point of Surveillance Capitalism is to perfect prediction machines based on all kinds of behavioral excess. So, instead of perceiving the users as a Monolith, one could dissect them back into smaller communities based on the data allocation itself or, like in the air example, into communities of countries or cultures. This begs the question if maybe the data extracted on such a scale is simply not a commons as it is not linked to any reasonable form of community. Here, the idea of enclosure and the logic of accumulation presented by Zuboff come into play.

It is not a coincidence that the extraction imperative of Surveillance Capitalism is unfamiliar to most of those who sustain it. From the beginning, Zuboff traces a line of policies and business decisions that masked the actual value of the extracted behavioral surplus. Companies and entities deploying these algorithms benefit from unawareness regarding the potency of their predictive capabilities. For example, consider the 2008 US presidential campaign by Barack Obama, where they used predictive algorithms to help sway the people deemed indecisive before the election (Zuboff, 2019, 122ff.). The behavioral data was used highly effectively, without anyone communicating what was used and how these datasets were acquired. In this context, the commons provide an exciting concept prevalently used in economics: the enclosure movement.

The concept of the enclosure movement initially described the legal process of removing or changing communal rights over land for the benefit of the new owners of these plots of land (Kain et al., 2004; Polanyi, 2024). In a broader sense, the enclosure movement describes the transformation of common goods into some form of private property (or state-owned property in the context of monarchies, for example). The idea plays a vital role when discussing natural commons but has since evolved into the more general description of turning commons into private or corporate property (Bollier, 2014, 41-54). A famous example would be that Warner Brothers owned the rights to the song "Happy Birthday" (Bollier, 2014, 65). Enclosure is thus a way to negate commons, as it removes it from the reach of a community and often integrates it into the capitalistic market[13]. The previous example with the women in India depended on the fact that they refused to commodify their seeds and their knowledge. A more dire example would be how Nestlé wants to enclose resources that could (or, as some would argue, should) be considered commons, like groundwater in Pakistan (Winschewski, 2017). Karl Polanyi (1944) described the effect and methodology of enclosure in detail in his work *The Great Transformation: The Political and Economic Origins of Our Time*. While the terminology between commons, commoners, and common people is a bit blurred, Bollier reads Polanyi as arguing that people lived in commons-managed communities for centuries before the emerging logic of capitalism commodified aspects of nature and labor (Bollier, 2014, 44f.; 91), further building this link between enclosure and commons. Moving to the enclosure of knowledge commons, Nancy Kranich writes about how even though more people than ever have access to a computer, there is a steady increase in information that is being privatized or removed from the public sphere in other ways (Kranich, 2006).

[13] While there are commons that operate within the capitalistic sphere, they do so managed by the community as some commons cannot be sustained without some interaction with surrounding economy. In this case the reference is specifically to resources that are enclosed as commons and inserted into the capital market under private ownership. E.g. Nestlé taking ground water and selling its bottles with it (Bollier, 2014).

So, to summarize, enclosure generally describes the process of removing a resource (and often this is a commons) from its managing community and transferring it to a government or into private property. But how does this relate to the digital commons of datasets?

Zuboff argues that the benefactors of Surveillance Capitalism attempted to conceal their production mechanism for a long time. At the same time, companies like Google and Meta lobbied for looser regulations regarding private data (Brodkin, 2017), attempted to shield themselves against lawsuits retroactively (Dougherty, 2016), or accused lawmakers of hindering progress with their measures for self-driving car safety (Bradshaw, 2015). Zuboff highlights these points to show how these companies were inherently compelled to lawlessness as they tried to gain more control over data and their options with it (Zuboff, 2019, 105). However, in the context of digital commons, we can also see the process of enclosure.

As companies discovered behavioral excess and its resulting use cases, they privatized the result and the resource. They began trading and selling data, be that private or general behavioral data, effectively removing them from the reach of communities in two ways. They did not clarify the relation between the data and their creation, making most users unaware of them, and they treated the resulting resource as their resource (Zuboff, 2019). As governments were equally unaware of these processes and were generally behind in regulating the online space, there were no rules to stop these companies as the logic of Surveillance Capitalism dictates a global harvesting of behavioral excess that denied its creators any access (Zuboff, 2019, 103-107). However, can we speak of enclosure when the companies bear the extraction costs? A strong argument against framing this as enclosure would be to say that the datasets do not exist without the tech companies extracting them, therefore making it legitimate that they keep the resources they help create. However, this argument fails in two aspects and is questionable in a third. The first argument is one of asymmetrical information. The second concerns the means of production, and the third lies in the relational nature of how AI interprets data.

It is a common problem in economics that if someone does not possess all the information, they cannot fix their price, wage accurately, and so on. The majority of users are unaware that they produce a resource leading to a public discourse that is centered around personal data and not around behavioral surplus data. This argument is even stronger when considering the forms of behavioral modification that are being conducted based on the data collected by companies. Moreover, as we have seen, this asymmetry is created and upheld purposefully. When we look at Public Open-Source data, we face a similar asymmetry. The people behind Wikipedia could not have anticipated companies like OpenAI, thus not shielding the commons of Wikipedia against them. But suddenly, various private entities abused that their most needed resource is for free. Using the commons to perpetuate their capitalistic endeavor.

The second argument would be one about the means of production. As Zuboff highlighted, the discovery of behavioral access was an accident. It is not needed to fulfill any of the actual goals of the various tech giants. It is neither inherently needed for the Google Search algorithm nor Facebook. Initially, the data was used solely to enhance the product experience. Make Google more accurate, help Facebook find your friends quicker, and so on (Zuboff, 2019, 70f.). But this new behavioral access is to a) alter your behavior and b) make you predictable for an algorithm. This access data is the crucial resource that is being harvested. From this perspective, the companies need two things: 1) a platform that enables the display and generation of behavioral excess, and 2) a platform that enables the display of commercials. In YouTube, for example, all three things are satisfied with a person visiting the platform, as the interaction with the algorithm provides behavioral surplus (1) while the videos are prefaced with a commercial (2). The trade deal is not that one gets to enjoy entertaining videos and, in return the company gets my data, but rather the platform infrastructure is the needed baseline to uphold this business and is the tool to enable behavior. It is also the platform to display the result of the behavioral access. Continuing this analogy, it would be as if we would argue that someone producing something in a factory should not be compensated because they enjoy working with the tools provided. Zuboff argues that we are not workers in Surveillance Capitalism as we are not being paid, but I would argue that makes us merely exploited workers. But in one sense or the other we are working/producing for big tech companies by providing the behavioral surplus as well as the training grounds for their AI. To summarize, the relationship between users and companies is an obscured one. It is not a straightforward transaction of labor and resources, as the users are unaware of what they are creating. While it is true that companies face the cost of collecting and storing the data, there appears to be no inherent reason to ascribe the rights to them.

The third argument is a relational one. The second argument could be used to claim that it still does not qualify as an enclosure of commons, as the working argument applies to singular agents. But the context of behavioral excess is a relational one. We react and behave in the context of other users, websites, and other creators on the internet. When we Google something, we look up other websites or digital commons. The recorded behavior is most useful as an observation of a person being a part of a web of human interactions. We can see this because behavioral data is not collected by just a few people or the same person repeatedly. It is collected from all of us to find common denominators within our interactions. It is tempting to perceive the machines of Surveillance Capitalism as algorithms that occupy themselves with singular persons and then mash together all the data it has. However, once the algorithm extracts the data, it plays a role in much more than just predicting us. Pooling together factors like approximate location, time of search, and search content, data can reveal your in-groups without collecting personal information and as we have seen in Chapter II.II AI wants to find the connecting dots, so the data one helped to create will help in understanding another connecting line altogether. Remembering the MNIST data set, the AI has no inherent understanding of a 2, and it needs to learn the difference in demarcation to other numbers. Similarly, AI has no conception of persons in the same sense that we do and needs to demarcate our data from the data of other persons. Large Language Models like ChatGTP need data pieces that contain texts, words, or messages to imitate humane language and knowledge, and it is the relative structure between them that enables the machine to mimic us (*How ChatGPT and Our Language Models Are Developed | OpenAI Help Center*, n.d.). Just as these algorithms are not treating the collected data as a collection of singular instances, these data sets can be framed as collaborative efforts that display human interaction and relational behavior with each other.

Finally, there is the aspect of ownership. It seems intuitive to assume that to train an AI, the respective programmer needs to "own" the data or at least have it stored in some way. Looking at ChatGTP, the company writes that "It does not 'copy and paste' training information – much like a person who has read a book and sets it down, our models do not have access to training information after they have learned from it" (*How ChatGPT and Our Language Models Are Developed | OpenAI Help Center*, n.d.). If that is accurate, one can see how the discourse on data privacy fails to capture the consequences of AI. The AI can learn by running the data in their training, without a company ever having true access to them. It is noticeable how OpenAI tries to convey this image of a company that is leaving your personal data alone (ibid). Only under the lens of Surveillance Capitalism can we see this second layer of data that concerns us and does not need to be saved somewhere in order to fulfill its function.

III. III A SET OF RULES FOR DATA

Now to the last aspect of our digital commons, the set of rules. This is the aspect most dependent on the other two. There can be a resource and a community without having specific rules[14], but the rules only exist concerning the community and the resource. Bollier writes: “There is no master inventory of commons because a commons arises whenever a given community decides it wishes to manage a resource in a collective manner, with special regard for equitable access, use, and sustainability.“ (Bollier, 2016, 175). So far, there is a resource and a loose conception of a potential community. However, without defining the community, it is hard to envision a set of rules to manage these commons. Depending on the definition of communities, the range of feasible options varies from national legislative orders to smaller communities organizing themselves or utilizing software that interferes with the processes of surveillance capitalism.

Looking at natural or knowledge commons, we face commons that existed long before enclosure or capitalism emerged. With behavioral surplus and similar data, we face the situation that this commodity never had the chance to be a commons. Thus, no set of social rules was established for it. Where most natural resources, like seeds in India, had previous rules, knowledge, and traditions that only needed to be rediscovered, these new data commons cannot refer to such methods. If we look at the discourse of trying to manage data via privacy concerns, one might argue that this shows how we must link data with something familiar, as it is a newer technology. A counterargument to that might be that we had data for a very long time, so this is not something new. But comparing data and statistics of the 20th century with the complex machinery described by Zuboff seems naïve at best and dangerous at worst. While it is well known that data and statistics have a long history of causing problems by either accidental or intended wrong interpretation (O’Neil, 2016), they have not been their industry with their logic of accumulating data and profit.

Turning to the open-source data that has not been created by Google, we can see a similar problem. These Open-Source datasets, like Wikipedia or Project Gutenberg, were presumably not created in anticipation of ever being a valuable resource for specific actors. As data consumption is often non-rivalrous, it can go unnoticed regarding how much private companies are profiting from them, as a company can use the entirety of Wikipedia for training

---

[14] In this case we would simply not talk about the relation as commons but the resource itself as well as the community would still exist.

their AI without anyone losing information. The problem here is not that others are profiting from these open-source platforms, after all, that was their key idea, but rather that this data can be used for the product of AI, something that people might object to. With AI applications reaching from projects like researchers using Large Language Models to decipher animal noises and construct "animal languages" (Bushwick, 2023) to the problematic prediction machines that Zuboff discusses, the rules governing these commons might need to be reformulated. OpenAI, the company behind ChatGTP, for example, notes on their website that their data comes from three sources: 1) publicly available data on the internet, 2) licensed third-party content, 3) the data obtained by their own platform through users (*How ChatGPT and Our Language Models Are Developed | OpenAI Help Center*, n.d.). In the first case, the previous argument applies already. Regarding point two, it is likely that these licensed data sets contain other data obtained via Surveillance Capitalism, but this intuition is still in need of proving by more rigorous research in this area. The third source of data brings us back to the YouTube argument postulated in Chapter III.II. The users of ChatGTP may have a skewed perception of what the trade off is. Because OpenAI needs people to interact with their AI to further improve it, meaning the users supply feedback and data. The argument at hand is not that people will oppose the usage nor that it is inherently problematic, but rather that the question of whether AI should be allowed to train on certain data is a discussion currently not held with the people who created the data.

The lack of a transparent community, combined with a widespread unawareness of the mechanisms behind this new logic of accumulation, led to a potential commons without a set of clear rules. The logic of surveillance capitalism performs a soft enclosure on these potential commons, removing a chance to treat them as such while they are created. Furthermore, digital commons face the predicament of often having no clear set of rules associated with them, as the process of enclosure happened too fast or was there to begin with. This paradox seems new and unique to these potential digital commons, as all other commons evolved as such and thus developed a set of rules for being governed. Now, we are faced with a backward operating enclosure movement that denies the status of the common before it could ever achieve it.

IV Conclusion

So, can the datasets used for AI be considered digital commons? Per definition, our answer at this time must be no. While I managed to show that data can be considered a resource, I failed to find any evident community and thus failed to find a set of rules regarding the governance of the resource. I displayed how this current status quo aligns with the logic of what Shoshanna Zuboff defines as Surveillance Capitalism. I showed that there are several reasons to frame these datasets as digital commons or at least as commodities capable of being commons but that the current understanding of labor and production on the internet conveys a distorted image of data that makes it harder to see data as such. Approaching the topic from a different angle than privacy, the paper has provided a discussion point that encapsulates much more data and highlights that personal data is only one kind of data that can heavily impact us and thus should concern us. Focusing on the data used for AI, the paper has elaborated on the need to gain a clearer understanding of how online data is acquired and produced. While the focus on the commons has opened a path that is less dependent on governmental regulation, it showed that some digital commons seem nearly impossible to manage. The relational nature of digital data provided a reasonable basis for the arguments for data as digital commons. However, it fell short of presenting any meaningful way of defining a community that cares for such things.

This thesis has ended in a position that enables several further research projects. From an economic perspective, it would be interesting to investigate the question of data production further. How do classical perceptions of labor, workforce, and so on, hold up in these digital spheres? Other interesting paths could be to apply models of game theory to the role of a user to determine the best response behavior once someone knows about the mechanism of surveillance capitalism, especially as Ostrom, for example, used game theory elements like the Prisoners Dilemma to argue for her points regarding the commons and our misconceptions regarding them (Ostrom, 2015).

The subject of this paper is a bit of a hybrid. Zuboff's Surveillance Capitalism is uncentered and scattered across various fields, combined with the activist approach of Bollier, this thesis’s basing lacks hard-scientific grounding. Aspects like AI programming, dataset analysis, or the nature of data need to be cross-checked by computer scientists, while the economic dimension needs proper modeling or mathematical expression in microeconomics to productively continue working on this subject. For now, I argue that this paper provides a productive theoretical exploration of the relation between data, commons, and AI that can serve as a starting point for others. I presented arguments that question current beliefs about our interactions on the internet, as well as explorations of the problems with potential commons in the digital sphere. Especially in the context of enclosure and digital data, the ethical dimension needs to be explored further.

It is nearly impossible for us, as users of the digital sphere, to negotiate the usage of data if we do not realize what role we play in the large chain of production. It is impossible to manage or even see the commons if one's perception is that of a loose link in a vast digital ocean. The data produced on the internet can be used for marvelous tools or heinous destruction. The same core idea can lead to an AI scanning X-rays for cancer or an AI seeking targets for a missile. As those who contributed a large part of the data to create these algorithms, there is a strong intuition that the decision of what to use data for should not belong to single persons, companies, or states. The commons provide an intriguing approach to governing data ourselves. But for this, we first need to realize data as something that could become commons. Thus, if further research proves the arguments of this thesis right, then we might need a new update to our common(s) knowledge.

V Sources:

V.I Literature Sources:

About CC Licenses. (2019). Creative Commons. Retrieved on 11 September 2024 from https://creativecommons.org/share-your-work/cclicenses/

About the Commons. (n.d.). The International Association for the Study of the Commons. Retrieved 18 September 2024, from

Artificial intelligence Definition & Meaning—Merriam-Webster. (n.d.). Retrieved 13 September 2024, from https://www.merriam-webster.com/dictionary/artificial%20intelligence

Bauwens, M., Kostakis, V., Tallinn University of Technology; P2P Lab, EE, Pazaitis, A., & Tallinn University of Technology; P2P Lab, EE. (2019). *Peer to Peer: The Commons Manifesto*. University of Westminster Press. https://doi.org/10.16997/book33

Berlinguer, M. (2020). NEW COMMONS: TOWARDS A NECESSARY REAPPRAISAL. *Popular Communication*, *18*(3), 201–215. https://doi.org/10.1080/15405702.2020.1781857

Birkinbine, B. J. (2020). *Incorporating the digital commons: Corporate involvement in free and open source software*. University of Westminster Press.

Bollier, D. (2006). The Growth of the Commons Paradigm. In C. Hess & E. Ostrom (Eds.), *Understanding Knowledge as a Commons* (pp. 27–40). The MIT Press. https://doi.org/10.7551/mitpress/6980.003.0004

Bollier, D. (2014). *Think like a commoner: A short introduction to the life of the commons*. New Society Publishers.

Bradshaw, T. (2015, December 18). *Google hits out at self-driving car rules*. https://www.ft.com/content/d4afee02-a517-11e5-97e1-a754d5d9538c

Brodkin, J. (2017) *Google and Facebook lobbyists try to stop new online privacy protections | Ars Technica*. Retrieved 17 June 2024, from https://arstechnica.com/tech-policy/2017/05/google-and-facebook-lobbyists-try-to-stop-new-online-privacy-protections/

Broussard, M. (2019). *Artificial unintelligence: How computers misunderstand the world* (First MIT Press paperback edition). The MIT Press.

Bushwick, S. (2023, May 1). *How Scientists Are Using AI to Talk to Animals*. Scientific American. https://www-scientificamerican-com.uaccess.univie.ac.at/article/how-scientists-are-using-ai-to-talk-to-animals/

Calzada, I., & Almirall, E. (2020). Data ecosystems for protecting European citizens' digital rights. *Transforming Government: People, Process and Policy*, *14*(2), 133–147. https://doi.org/10.1108/TG-03-2020-0047

Calzada, I., Pérez-Batlle, M., & Batlle-Montserrat, J. (2023). People-Centered Smart Cities: An exploratory action research on the Cities' Coalition for Digital Rights. *Journal of Urban Affairs*, *45*(9), 1537–1562. https://doi.org/10.1080/07352166.2021.1994861

Carballa Smichowski, B. (2016). *Data as a common in the sharing economy: A general policy proposal*.

Carballa Smichowski, B. (2019). Alternative Data Governance Models: Moving Beyond One-Size-Fits-All Solutions. *Intereconomics*, *54*(4), 222–227. https://doi.org/10.1007/s10272-019-0828-x

Cinnamon, J. (2017). Social Injustice in Surveillance Capitalism. *Surveillance & Society*, *15*(5), 609–625. https://doi.org/10.24908/ss.v15i5.6433

Cohen, Julie E. 2019. Review of Zuboff's *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. *Surveillance & Society* 17(1/2): 240-245. http://library.queensu.ca/ojs/index.php/surveillance-and-society/index *commons, n. Meanings, etymology and more | Oxford English Dictionary*. (n.d.). Retrieved 25 May 2024, from https://www.oed.com/dictionary/commons_n

Community Definition & Meaning—Merriam-Webster. (n.d.). Retrieved 21 September 2024, from https://www.merriam-webster.com/dictionary/community

Creemers, R. (2018). China's Social Credit System: An Evolving Practice of Control. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3175792

Cuquet, M., Vega-Gorgojo, G., Lammerant, H., Finn, R., & Hassan, U. ul. (2017). *Societal impacts of big data: Challenges and opportunities in Europe* (arXiv:1704.03361). arXiv. http://arxiv.org/abs/1704.03361

DATA | Bedeutung im Cambridge Englisch Wörterbuch. (n.d.). Retrieved 6 June 2024, from https://dictionary.cambridge.org/de/worterbuch/englisch/data

Dougherty, C. (2016, May 28). Tech Companies Take Their Legislative Concerns to the States. *The New York Times*. https://www.nytimes.com/2016/05/28/technology/tech-companies-take-their-legislative-concerns-to-the-states.html

Fia, T. (2021). An Alternative to Data Ownership: Managing Access to Non-Personal Data through the Commons. *Global Jurist*, *21*(1), 181–210. https://doi.org/10.1515/gj-2020-0034

Fuchs, C. (2021). The Digital Commons and the Digital Public Sphere How to Advance Digital Democracy Today. *Westminster Papers in Communication and Culture*, *16*(1). https://doi.org/10.16997/wpcc.917

Garcia-Martinez, A. (2017, May 2). I'm an ex-Facebook exec: Don't believe what they tell you about ads. *The Guardian*. https://www.theguardian.com/technology/2017/may/02/facebook-executive-advertising-data-comment

Garrett, P. M. (2022). 'Surveillance Capitalism, COVID-19 and Social Work': A Note on Uncertain Future(s). *The British Journal of Social Work*, *52*(3), 1747–1764. https://doi.org/10.1093/bjsw/bcab099

Grossman, R. L., Heath, A., Murphy, M., Patterson, M., & Wells, W. (2016). A Case for Data Commons: Toward Data Science as a Service. *Computing in Science & Engineering*, *18*(5), 10–20. https://doi.org/10.1109/MCSE.2016.92

Guha, R. V., Radhakrishnan, P., Xu, B., Sun, W., Au, C., Tirumali, A., Amjad, M. J., Piekos, S., Diaz, N., Chen, J., Wu, J., Ramaswami, P., & Manyika, J. (2023). *Data Commons* (arXiv:2309.13054). arXiv. http://arxiv.org/abs/2309.13054

Guo, E. (2022, December 12). *A Roomba recorded a woman on the toilet. How did screenshots end up on Facebook?* MIT Technology Review. https://www.technologyreview.com/2022/12/19/1065306/roomba-irobot-robot-vacuums-artificial-intelligence-training-data-privacy/

Haggart, B. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, S. Zuboff (2018). *Journal of Digital Media & Policy*, *10*(2), 229–243. https://doi.org/10.1386/jdmp.10.2.229_5

Hardin, G. (1968). The Tragedy of the Commons. *Science*, *162*(3859), 1243–1248.

Hardin, G. (*Tragedy of the Commons*. (n.d.). Econlib. Retrieved 26 May 2024, from https://www.econlib.org/library/Enc/TragedyoftheCommons.html

Havukainen, M., Mikkilä, M., & Kahiluoto, H. (2022). Global climate as a commons – Decision making on climate change in least developed countries. *Environmental Science & Policy*, *136*, 761–771. https://doi.org/10.1016/j.envsci.2022.08.003

Heath, A. P., Ferretti, V., Agrawal, S., An, M., Angelakos, J. C., Arya, R., Bajari, R., Baqar, B., Barnowski, J. H. B., Burt, J., Catton, A., Chan, B. F., Chu, F., Cullion, K., Davidsen, T., Do, P.-M., Dompierre, C., Ferguson, M. L., Fitzsimons, M. S., … Grossman, R. L.

(2021). The NCI Genomic Data Commons. *Nature Genetics*, *53*(3), 257–262. https://doi.org/10.1038/s41588-021-00791-5

Hess, C., & Ostrom, E. (2006). Introduction: An Overview of the Knowledge Commons. In C. Hess & E. Ostrom (Eds.), *Understanding Knowledge as a Commons* (pp. 3–26). The MIT Press. https://doi.org/10.7551/mitpress/6980.003.0003 and

A Framework for Analyzing the Knowledge Commons. In C. Hess & E. Ostrom (Eds.), *Understanding Knowledge as a Commons* (pp. 41–82). The MIT Press. https://doi.org/10.7551/mitpress/6980.003.0005

Hicks, J. (2023). The future of data ownership: An uncommon research agenda. *The Sociological Review*, *71*(3), 544–560. https://doi.org/10.1177/00380261221088120

History of Linux. (2023, February 12). GeeksforGeeks. Retrieved 12. August 2024 from

How ChatGPT and our language models are developed | OpenAI Help Center. (n.d.). Retrieved 21 September 2024, from https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

https://www.geeksforgeeks.org/linux-history/

Hyde, Lewis. *Common as Air: Revolution, Art and Ownership.* Farrar, Straus and Giroux, 2010.

Jaen, N. (2024, April 8). *Determining what AI data you need and how to source it*. Retrieved 29 July 2024, from https://www.rws.com/artificial-intelligence/train-ai-data-services/blog/determining-ai-data-needs-and-sourcing/

Jansen, S. C., & Pooley, J. (2021). Blurring genres and violating guild norms: A review of reviews of *The Age of Surveillance Capitalism*. *New Media & Society*, *23*(9), 2839–2851. https://doi.org/10.1177/14614448211019021

Jeong, J. C., Hands, I., Kolesar, J. M., Rao, M., Davis, B., Dobyns, Y., Hurt-Mueller, J., Levens, J., Gregory, J., Williams, J., Witt, L., Kim, E. M., Burton, C., Elbiheary, A. A., Chang, M., & Durbin, E. B. (2022). Local data commons: The sleeping beauty in the community of data commons. *BMC Bioinformatics*, *23*(S12), 386. https://doi.org/10.1186/s12859-022-04922-5

Johns, F. (2020). "Surveillance capitalism" and the angst of the petit sovereign. *The British Journal of Sociology*, *71*(5), 1049–1051. https://doi.org/10.1111/1468-4446.12784

Joyce, S. (2024) Surveillance Capitalism and its Subjectivity. *Political Perspectives*, *11*(1), 5-33.

Kain, R. J. P., Chapman, J., & Oliver, R. R. (2004). *The enclosure maps of England and Wales, 1595-1918*. Cambridge University Press.

Klein, N. (2008). *The shock doctrine: The rise of disaster capitalism*. Penguin Books.

Kranich, N. (2006). Countering Enclosure: Reclaiming the Knowledge Commons. In C. Hess & E. Ostrom (Eds.), *Understanding Knowledge as a Commons* (pp. 85–122). The MIT Press. https://doi.org/10.7551/mitpress/6980.003.0007

Laerhoven, F. V., & Ostrom, E. (2007). Traditions and Trends in the Study of the Commons. *International Journal of the Commons*, *1*(1), 3. https://doi.org/10.18352/ijc.76

LeCun, Y., Cortes, C., & Burges, C. J. C. (n.d.). *MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges*. Retrieved 01 September 2024, from https://yann.lecun.com/exdb/mnist/index.html

Lloyd, W. F. (1833). *Two lectures on the checks to population*. http://archive.org/details/twolecturesonch00lloygoog

Madison, M. J., Frischmann, B. M., & Strandburg, K. J. (2017). *Constructing Commons in the Cultural Environment*. https://doi.org/10.31228/osf.io/76g93

Marella, M. R. (2017). The Commons as a Legal Concept. *Law and Critique*, *28*(1), 61–86. https://doi.org/10.1007/s10978-016-9193-0

Mazé, A. (2023). Geographical indications as global knowledge commons: Ostrom's law on common intellectual property and collective action. *Journal of Institutional Economics*, *19*(4), 494–510. https://doi.org/10.1017/S1744137423000036

Morozov, E. (2019, February 4). *Capitalism's New Clothes*. The Baffler. https://thebaffler.com/latest/capitalisms-new-clothes-morozov

Naughton, J. (2019, January 20). 'The goal is to automate us': Welcome to the age of surveillance capitalism. *The Observer*. https://www.theguardian.com/technology/2019/jan/20/shoshana-zuboff-age-of-surveillance-capitalism-google-facebook

Neri, A. (2020, March 2nd). *We should treat data as a natural resource. Here's why*. World Economic Forum. https://www.weforum.org/agenda/2020/03/we-should-treat-data-as-a-natural-resource-heres-why/

Noack, P., & Sanner, S. (2023). *Künstliche Intelligenz verstehen: Eine spielerische Einführung* (2. aktualisierte Auflage). Rheinwerk Verlag.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Penguin Books.

Omoyiola, B. O. (2023). The social implications, risks, challenges and opportunities of big data. *Emerald Open Research*, *1*(4). https://doi.org/10.1108/EOR-04-2023-0014

Ostrom, E. (2015). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.

Papadimitropoulos, V. (2020). *The Commons: Economic Alternatives in the Digital Age*. University of Westminster Press. https://doi.org/10.16997/book46

Parthasarathi, P. (2024, June 13). *Narrow AI - Bannari Amman Institute of Technology*. https://www.bitsathy.ac.in/blog/narrow-ai/

Piekos, S. N., Gaddam, S., Bhardwaj, P., Radhakrishnan, P., Guha, R. V., & Oro, A. E. (2021). Biomedical Data Commons (BMDC) prioritizes B-lymphocyte non-coding genetic variants in Type 1 Diabetes. *PLOS Computational Biology*, *17*(9), e1009382. https://doi.org/10.1371/journal.pcbi.1009382

Piketty, T. (2016). *Das Kapital im 21. Jahrhundert* (I. Utz & S. Lorenzer, Trans.; 1. Auflage in C.H. Beck Paperback). C.H. Beck.

Polanyi, K. (2024). *The Great Transformation: The Political and Economic Origins of our Time*. Penguin Modern Classics Series, Penguin Random House

Rosengrün, S. (2021). *Künstliche Intelligenz zur Einführung*. Junius.

Schwartz, P. (2020). *Climate Change: A Tragedy of the Commons?* Econlib. https://www.econlib.org/library/Columns/y2020/Schwartzclimatechange.html

Šestáková, A., & Plichtová, J. (2019). Contemporary commons: Sharing and managing common-pool resources in the 21st century. *Human Affairs*, *29*(1), 74–86. https://doi.org/10.1515/humaff-2019-0007

Siegmann, A. (2022). *Climate of the Middle: Understanding Climate Change as a Common Challenge* (1st ed. 2022.). https://doi.org/10.1007/978-3-030-85322-8

Smith, M. (2019, February 12th) *Data is the world's new natural resource*. IBM A/NZ Blog. https://www.ibm.com/blogs/ibm-anz/data-is-the-worlds-new-natural-resource/

Stobierski, T. (2020, November 12). *What Are Network Effects?* | HBS Online. Business Insights Blog. https://online.hbs.edu/blog/post/what-are-network-effects

Tang, J., Tanoli, Z.-R., Ravikumar, B., Alam, Z., Rebane, A., Vähä-Koskela, M., Peddinti, G., Van Adrichem, A. J., Wakkinen, J., Jaiswal, A., Karjalainen, E., Gautam, P., He, L., Parri, E., Khan, S., Gupta, A., Ali, M., Yetukuri, L., Gustavsson, A.-L., … Aittokallio, T. (2018). Drug Target Commons: A Community Effort to Build a Consensus Knowledge

Base for Drug-Target Interactions. *Cell Chemical Biology*, *25*(2), 224-229.e2. https://doi.org/10.1016/j.chembiol.2017.11.009

Van Maanen, G., Ducuing, C., & Fia, T. (2024). Data commons. *Internet Policy Review*, *13*(2). https://doi.org/10.14763/2024.2.1748

Varoufakis, Y. (2023). *Technofeudalism: What killed capitalism*. The Bodley Head.

What Is Artificial Intelligence (AI)? | IBM. (2024, August 9). https://www.ibm.com/topics/artificial-intelligence

Whitehead, M. (2019). The Age of Surveillance Capitalism - A Review Essay. *Antipode*, 1.

Winschewski, J. (2017). Putting a Price Tag on Human Rights. An Anthropological Perspective on Nestle's Drinking Water Privatisation in Pakistan. *Prace Etnograficzne*, *45*(2). https://doi.org/10.4467/22999558.PE.17.009.7904

Yakowitz Bambauer, Jane R., Tragedy of the Data Commons (March 18, 2011). Harvard Journal of Law and Technology, Vol. 25, 2011, Available at SSRN: https://ssrn.com/abstract=1789749 or http://dx.doi.org/10.2139/ssrn.1789749

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile books.

Zygmuntowski, J. J., Zoboli, L., & Nemitz, P. F. (2021). Embedding European values in data governance: A case for public data commons. *Internet Policy Review*, *10*(3). https://doi.org/10.14763/2021.3.1572

V.II Image Sources:

Image 1;

3.3. The MNIST Dataset—Conx 3.7.9 documentation. (n.d.). Retrieved 23 September 2024, from https://conx.readthedocs.io/en/latest/MNIST.html

As the MNIST dataset is publicly available, no Copyright Claims to displayed numbers from the dataset were found.

V.III Video Sources:

Bloomberg Technology. (2024, May 20). *Nvidia CEO Says Data Is a Natural Resource*. Retrieved on 7. August 2024 from https://www.bloomberg.com/news/videos/2024-05-20/nvidia-ceo-says-data-is-a-natural-resource-video

Sanderson, G. (Director). (2017, October 5). *Aber was ist ein neuronales Netzwerk? | Kapitel 1, Deep Learning* [Video recording]. https://www.youtube.com/watch?v=aircAruvnKk

V.I Abbreviations:

AI: Artificial Intelligence

CPR: Common Pool Ressource

KI: Künstliche Intelligenz

LLM : Large Language Modell

V.II Appendix

Abstract English

This thesis asks if we can perceive datasets used to train AI as digital commons. For this, it uses the definition of the social activist and scholar David Bollier, who defines commons as the relational combination of a resource, a community, and a set of rules (2014). To limit the scope of the paper, it takes two examples of dataset-AI-relations: a) Wikipedia as an Open-Source Commons and the Large Language Model like ChatGTP and b) the narrow AI of the Google Search algorithm and the data that the philosopher and economist Shoshanna Zuboff calls behavioral access. Zuboff's work, *The Age of Surveillance Capitalism* (2019) provides an economical-philosophical framework that introduces a specific logic of data accumulation focused on collecting data to perfect algorithms that predict people or a group of people and exert power via this mechanism. The conclusion of this thesis is that the discussed datasets cannot be seen as digital commons under the proposed framework because neither the aspect of community nor the aspect of a set of rules meet the criteria of the commons definition to a satisfactory degree. However, the paper also highlights how these circumstances could be intended consequences of the inherent logic of Surveillance Capitalism. Thus, the thesis highlights the tension between an intuitive viewing of these data sets as commons and the reasons for obstructing this framing.

Abstract Deutsch / German

Diese Arbeit stellt die Frage ob es möglich ist Daten, welche für das Training von Künstlicher Intelligenz (KI) benützt wird, als Digitales Allmendegut zu sehen. Dafür wurde die Allmendegut Definition von David Bollier hergenommen, welche ein Allmendegut als eine relationale Beziehung zwischen einer Ressource, einer Gemeinschaft und einem Regelwert sieht (2014). Um das Ausmaß der Arbeit zu begrenzen, bezieht sie sich auf die Datensätze, die für den Googel Suchalgorithmus benützt werden, sowie die Open-Source Daten wie Wikipedia, die ChatGTP benützt. Außerdem wir die Arbeit *The Age of Surveillance Capitalism* (2019) von Shoshana Zuboff herbeigezogen, um einen wirtschaftlich-philosophischen Rahmen zu bieten, welcher sich mit der von ihr beschriebene Logik der Daten Akkumulation, zur Berechnung von Menschen beschäftigt. Die Konklusion der Arbeit ist es, dass wir nicht von Digitalen Allmendegütern sprechen können, da die Aspekte von Gemeinschaft und Regelwerk nicht zufriedenstellen erfüllt werden. Jedoch stellt die Arbeit fest, dass es hierfür Gründe gibt, die sich erst im Rahmen von Zuboff‘s Überwachungskapitalismus offenbaren. Dadurch bietet die Arbeit mehr Kontext für die Spannung zwischen der intuitiven Einordnung dieser Datensets als Allmendegüter und den Gründen die eine solche Einordnung erschweren.