Data pollution: Definition and Policy Responses

Leonardo Mori[1][0009-0006-6305-6564], Alizée Francey[1][0000-0001-8832-1332] and Tobias Mettler[1] [0000-0002-7895-7545]

[1] Swiss Graduate School of Public Administration, University of Lausanne, Chavannes-près-Renens, Switzerland

leonardojacopo.mori@unil.ch

Abstract. Data has become one of the strongest drivers of econ
omic growth and innovation. However, this data-driven transformation brings various challenges and harms affecting our lives and environments across different domains and scales. In this article, we define the set of such harms as *data pollution*. Data pollution is a multifaceted phenomenon, entailing different dimensions and complex mechanisms, which we capture in one conceptual model using network thinking and cybernetics. We further analyse the policy landscape to comprehend the awareness level and responses to this phenomenon.

Keywords: Data pollution, conceptualisation, model, cybernetics.

1 Introduction

The surge in data, often termed the *data deluge*, is a direct outcome of substantial advancements in various domains, including the widespread adoption of smartphones, the ubiquity of social media platforms, the pervasive integration of the Internet of Things (IoT), among others [1,2]. These global trends generate more data than ever, and by 2025, there will be more than 175 zettabytes of data, reflecting a fivefold growth from 2018 to 2025 [3,4]. This significant growth in data volume and the rates at which data are generated make data the lifeblood of the economy and a driver of innovation and societal progress, notably through the progressive extension of Artificial Intelligence (AI) use, which itself necessitates the analysis of extensive volumes of data [5].

Within this context, the *data economy* emerged as a catch-all term covering all aspects related to the generation, collection, storage, processing, sharing, analysis, and use of data facilitated by digital technologies [4]. Data are thus considered a valuable asset, as most economic activities may depend on data within a few years [5]. While the value of the data economy of EU27 was almost €325 billion in 2019, representing 2.6% of the gross domestic product (GDP), predictions foresee that the European Union (EU) data economy will be worth €550 to €829 billion in 2025, representing 4% to 6% of the overall EU GDP [4,5]. To achieve this, the European Commission supports data sharing through legislation and practical measures, notably by publishing a sequence of directives, strategies, and regulatory acts to set directions for the EU member states [4]. As illustrative examples, critical documents include the General Data Protection Regulation (GDPR) (2018), which sets standards for data privacy; the Open Data Directive (2019), which provides standard rules for government-held data by addressing barriers to the reuse of publicly funded information; the European Data Strategy (2020), which is oriented toward establishing a unified data market that not only bolsters Europe's competitiveness but also fortifies its control over the data; the Data Governance Act (2021), which promotes the availability of data by allowing reuse of some categories of protected public sector data; and the Data Act (2022), which sets rules for the use of data generated by IoT-enabled devices [4]. Moreover, the creation of nine European data spaces aims to facilitate the secure and cost-effective exchange of data across the EU, encompassing both public sector and business data to stimulate the growth of novel data-driven products and services [5,6]. In addition to promoting data sharing, government at all levels and from all parts of the World have been dedicating remarkable effort to the digitalization of their own operations for now decades [7,8], themselves largely contributing to the data deluge [9]. This trend is bound to accelerate in the coming years, as a 2022 survey to senior officials of 200 city governments across the World reports that respectively 73% and 49% of respondents identified making real-time decisions from data and making data accessible to the public as priorities for the 5 years to come [10].

As data fuels the new economy by creating endless opportunities, it becomes to this century what oil was to the last one, and it thus pollutes [2,11,12]. As an illustrative example, the Shift project, a think tank promoting the transition to a post-carbon economy, has estimated that the proportion of worldwide greenhouse gas emissions attributable to data has risen from 2.5% in 2013 to 3.7% in 2019 [12,13]. However, harms generated by data expand well over ecological pollution, and following the pivotal role of data in the new economy, this paper argues for the imperative to acknowledge such concomitant deleterious harm; we refer to these as *data pollution*. Since the 1980s, data pollution has been used in different contexts with different meanings, including insufficient quality data [14,15] but also the external repercussions that emanate from data use such as social and ecological side effects [11,12]. We argue that these different meanings all emanate from a common cause, being the massive increase in data availability, and interact with each other through cause to effect relationships and ultimately forming a complex problem. This article thus does not propose a new definition of data pollution per se, but rather to relate the different existing meanings through a conceptual model dissecting the dynamics contributing to data pollution. This article defines data pollution as the set of harms generated by any data activity. The conceptual model's primary objective is to develop a comprehensive understanding of data pollution to render these harms perceptible systematically and acknowledged while concurrently elucidating the existing regulatory mechanisms at the EU level. We believe that such an understanding is particularly relevant for governments, both for evaluating the impacts of their own data related initiatives, and to regulate those from the private sector. Finally, we conclude that addressing data pollution with a comprehensive perspective is necessary for the promotion of sustainability, responsible data management, and the protection of individual and societal well-being.

2 Background

2.1 Data pollution definitions in the extant literature

Data pollution has been used in various pieces of research since the 1980s. Over the decades, it has been employed by different authors from different research disciplines, such as machine learning [16], computer networking [17], AI research [18], and even neuropsychiatry [19]. These disciplines are highly diverse, and the meaning given to data pollution in extant literature is varied and strongly dependent on the context in which it is used, sometimes referring to pollution *of* the data, and other times to pollution *by* the data. Nonetheless, data pollution is often used to designate recurring phenomena with three frequently associated meanings.

The first mention of data pollution dates from 1986 and refers to it as "*the accumulation of all 'contaminations' or 'distortions' which can result from working with data in the information technology field*" ([15], p. 291). Still today, data pollution is often used to refer to data that is of bad quality, untrustworthy, or not predisposed to be used optimally. This broad category includes a spectrum of aspects associated with data pollution, which, although similar, often differ as they may be specific to particular contexts or problematics. Indeed, while older works would be concerned with a general "*contamination of the information supply with incomplete, inconsistent, or incorrect information*" ([14], p. 24), many authors have used the term data pollution to designate more field-specific problematics in recent years. For instance, researchers in machine learning tend to include sample imbalances within data pollution [18,20]. Another way the meaning of data pollution has been restricted is by referring to it as introducing inaccurate or otherwise unhelpful data into datasets rather than to the exitance of data. Along these lines, publications about network coding systems almost exclusively mention data pollution in the context of attacks, in which "*attackers inject corrupted [data] packets into the network*" ([21], p. 741), while other scholars go as far as defining data pollution based on the unintentionality to introduce errors into the data, in contrast with data poisoning, which refers to voluntary data degradation [19]. In all these cases, polluted data are considered a nuisance because they potentially have adverse effects on the performance of their intended use.

A second meaning of data pollution relates to excessive data production, storage, and publication. Here, data are considered to pollute by their mere existence, as they do not generate benefits and only take up storage space [22]. This happens when data are duplicated or disseminated without there being an interest in it. Not only are these datasets useless, but they can contribute to decreasing the findability of data we would like to use and to a series of unwelcomed consequences, such as privacy violation [22] or even negatively impact our capacity to recognise information from fake news, our concentration and overall well-being [23].

Lastly, recent discussions have emerged over the unwanted effects of data on social environments. According to Ben-Shahar [11], data pollution refers to the harmful effects of the "*exchange of data between giver and taker*" ([11], p. 148). The author argues that while private data leaks are often considered detrimental because of their harm to privacy, their potential damages go well beyond that, as he believes that data production "*creates public harms and destroys public goods*" ([11], p. 106).

2.2 Toward an overarching definition of data pollution

Although extant literature has identified several meanings and related aspects that can be attributed to data pollution, each was conceptualised and is usually considered in isolation from the others. However, we argue that the various meanings and related aspects of data pollution have a shared origin arising from the increased ability to produce, store and use vast amounts of data and the adjustments of businesses and public organisations' practices under the big data era and its associated data economy. There is thus a gap in the existing literature in understanding data pollution, given that there is currently no overarching view that would effectively contextualise its diverse meanings and related aspects, delineate their origins, and elucidate their complex interactions. Thus, we posit that the interactions between various notions existing in the scientific literature, or even the public debate, and data pollution should be investigated. In pursuit of this, we contend that in addition to those mentioned above, other issues resulting from data activities should be incorporated into the definition of data pollution.

The first of these aspects is the notion of data overload, which in the context of lexicography has been defined as a situation where "the dictionary user gets more data than he or she needs or can deal with during the present consultation and becomes confused and fails to retrieve the necessary information" ([24], p. 397). This idea relates to the over-publication of data mentioned above and the concept of information overload, defined as a situation where "information received becomes a hindrance rather than a help when the information is potentially useful" ([25], p. 249), in some extreme cases even leading to health issues. Thus, we believe that data overload should be included in an overarching definition of data pollution, as we consider it a set of harm resulting from the data. By including data overload, we acknowledge the social adverse effects of excessive exposure and the fact that maintaining and processing excessive volumes of data requires substantial infrastructures and resources, contributing to energy consumption.

The growing mindfulness of environmental challenges is spurring a debate over the ecological pollution related to data. It was estimated that in 2019, digital technology produced 4% of the overall greenhouse gas emissions [26], while data centres and transmission networks alone are the cause of 1% of energy-related greenhouse gas emissions, as a 2023 study by the International Energy Agency (IEA) revealed [27]. In 2022, global data centre electricity consumption accounted for 1 to 1.3% of global final electricity demand, with an annual growth of 20 to 40% in the latest years ([27], p. 2). Given the trend of continuing digitalisation of products and services, carbon emissions, natural resource extraction, production of waste, and other harmful environmental impacts, directly or indirectly, will raise with data-driven infrastructures gaining in economic importance [28]. Such waste is fuelled by the proliferation of electronic devices and their rapid obsolescence. Given that a proportion of the current ecological pollution is due to the set of harm from the data or its use, we believe that it should be integrated into the overarching definition of data pollution to highlight the environmental consequences of our digital lifestyle.

3 Methodology

Given that data pollution exhibits a multifaceted nature, involving diverse dimensions and mechanisms, our decision to utilize network thinking and cybernetics as an approach aims to enhance our understanding of this phenomenon. The central focus of cybernetics is not so much on the structural elements within the system but on their operational dynamics [29]. Cybernetics acknowledges that our understanding of systems relies on our simplified representations or models of those systems and also recognises that simplified representations or models ignore aspects of the system irrelevant to the purpose for which the model is constructed [30]. While data pollution itself is not a complex system, the context from which it stems can be characterised as complex due to the interactive components involved in the emergence of data pollution. We thus understand data pollution within a more extensive system that can be seen as complex. As their name suggests, complex systems are typically hard to understand, but network thinking may ease their comprehension [31]. We thus used network thinking to build our conceptual model on data pollution, which is a collection of nodes and links between these nodes. While the interactions between its different components make the definition and management of data pollution difficult, the feedback loops make detecting and remedying data pollution complex, as actions taken to prevent data pollution may have subsequent effects. This is especially true given that an effect can feed back into its cause in cybernetics. For example, many algorithms use data to propose products and services. If the data are somehow dirty, the algorithms may create bias and inequalities, which may generate dirty data feeding back to the data used by algorithms, leading to increased bias and inequalities, thus creating a vicious circle. Moreover, feedback loops can be positive or negative [30,32,33]. The feedback loop is negative if a positive deviation leads to a negative deviation at the following node. For instance, when there is a rise in the volume of data, it reduces search efficiency. Subsequently, this reduction in search efficiency leads to a decline in the retrievable data volume, as only a fraction of the data can be accurately found. The opposite situation, where an increase in the deviation produces further increases, is called a positive feedback loop. For example, a rise in the volume of data will allow for more innovative products and services that will themselves generate data, thus leading to higher data volume. This straightforward method allows us to ascertain whether a given loop will result in stabilisation (indicative of a negative feedback loop) or an unrestrained and escalating process (indicative of a positive feedback loop) [30].

Our research started with the identification of key nodes involved in the genesis and propagation of data pollution. This analytical foundation enabled us to construct a nuanced and systematic representation of the complex interactions shaping the landscape of data pollution. Altogether, the interactions between the nodes facilitated the creation of a visual representation of the system, which we refer to as the conceptual model. We then reviewed and refined our conceptual model by stimulating the analysis of different instances. This process significantly enhanced our conceptual model's precision and depth. Finally, we ensured that the conceptual model adeptly captured the nuances and dynamics of data pollution, which enabled the additional or removal of nodes, the recalibration of the interactions, and the incorporation of feedback loops. By incorporating the role of feedback loops from cybernetics, the model considers the interactions and interdependencies of the nodes. Moreover, in both network thinking and cybernetics, the flow of information is crucial and combining both shows how data pollution flows within and between the nodes. Combining network thinking and cybernetics can lead to a more comprehensive understanding of data pollution and provide valuable tools for managing and mitigating it. The objective of this model is to propose a new conceptualisation of the most discussed forms of data pollution, showing how they stem from a common phenomenon and interrelate. We hope that our model can stimulate discussion and serve as a basis for further work expanding it with other related data harms.

4 Conceptualising data pollution

Figure 1 shows our conceptual model, aiming to provide an overview of the complex mechanisms generating data pollution, which we describe in more detail next.

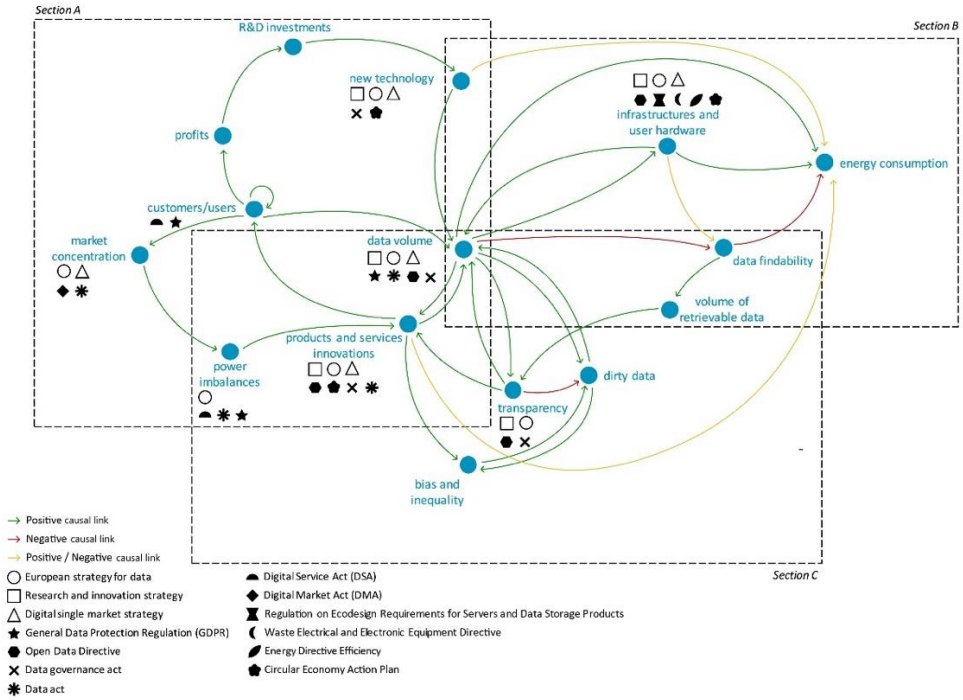

Fig.1. Conceptual model.

4.1 The data economy

Section A schematises the mechanisms feeding the pace of the data economy that we know today: the increased volume of data allows businesses to offer more tailored products and services, enhancing the customer experience they can deliver and consequently attracting more customers. This means that the volume of data businesses can collect equally surges alongside their profits. Increased profits translate into expanding opportunities for investment in further technological development, thus fostering the continuous growth of data. New technology bringing about new data collection possibilities is easily illustrated with IoT devices, which grant corporations access to new kinds of customer data [34]. It is crucial to emphasise that the different model nodes aren't always associated with a single organisation but rather encompass the broader data economy ecosystem. It is not necessarily the organisation registering the customer increase and its consequent growth in customer data and profits that will directly boost research and development (R&D) investments that will result in new technology. However, since one organisation will have more monetary capacity, it will feed the whole sector by making more orders from its suppliers and generally increasing the market size, allowing its other actors to boost their activities. As an illustrative example, we use smartphones, the data collecting devices, running the Android OS, which is arguably what generates the network effects mainly thanks to its Play Store. In this case, the hardware and the software are complementary products produced by different companies operating in an ecosystem with different roles. Indeed, smartphone manufacturers make the highest R&D investments and provide the technology necessary to collect certain data types. In contrast, the operating system developer provides the platform which arguably plays the biggest role in client attraction, as revealed by the failure of the Microsoft phone, which, despite proposing high-quality, innovative designs, ran on its own OS and struggled to attract third-party apps [35]. Even though the volume of data is at the model's centre, any node can be considered as a starting and ending point of the feedback loop. For example, concerning public sector services, one could argue that the surge in the volume of data comes from the digitalisation of services, itself induced by political will, rather than the development of revolutionary technology. In this case, the feedback loop would start with the product and services

innovation node. Alternatively, we could consider instances where the loop starts with increased customers. For example, such cases can happen with external factors influencing the market, as in several countries with governments subsidising electric vehicle purchases.

The extension of the left part of the section includes a highly simplified depiction of how the platform economy works. Indeed, the platform model has network externalities, which means that the perceived value offered by the platform increases as the number of users increases [36]. What characterises these platforms is that as the number of users grows, not only does the amount of data they can collect follows the same trend, but the platform will become increasingly attractive to other potential users and other market sides, which is a dynamic leading to market concentration and monopoly risks. Indeed, market concentration and the power it gives to large established companies allows them to acquire potential new market entrants, which could threaten their dominant position by eroding their user pool. This implies that companies that could change the status quo by bringing ethical solutions for users are often purchased by their competitors before they can grow enough to have an impact [37]. This concentration also results in power imbalances, with data subjects losing control over their data in favour of the data holders. This control loss can have consequences that data subjects would not expect and accept, such as transferring their data to third parties [38], prominently exemplified by the Cambridge Analytica scandal [11].

4.2 Data storage

Section B mainly concerns data storage. A diversity of infrastructures and user hardware are required to store and share data, including data centres, mobile networks, user equipment, and data portals, to mention a few examples. Naturally, as the volume of data grows, so does the need to expand the data storage and transfer capacity. This implies that the volume of infrastructures and hardware manufacturing would increase, and in step, the amount of electronic waste. However, the features and governance of the infrastructures can also affect the volume of data. The promoters of data spaces, for example, which have recently received much attention, argue that pooling data across the actors of strategic industries can give birth to innovative initiatives which none of the participating parties could have enacted by itself. These new initiatives typically generate data [39]. Furthermore, the quality of the infrastructure influences the findability of the data it hosts, which we argue is likely to decrease with the volume of data stored. This is because of the concept of data overload, introduced above, according to which having access to the immense volume of data will make it harder for potential data users to find precisely what they were looking for, and scarce findability will lead users to resort to search queries that require energy use [40]. More generally, data storage activities have ecological consequences. One of the primary contributors to the data storage carbon footprint is the energy consumed by data centres. According to the IEA [27], data centres' energy use is going to follow a growing trend in the years to come despite high-efficiency improvements which have allowed data centres energy consumption to increase at a rate between 20% and 70% globally while the number of data centre workloads recorded an inflation of 340% over the same period. All activities related to data-related hardware and its use inherently have environmental effects. This is observable in the polluting processes linked to resource extraction, manufacturing, and data transportation, and the resources essential for their functioning, such as cooling systems that demand substantial energy consumption.

4.3 Social implications

Our model's section C refers to the social implications related to the volume of data. As mentioned above, data are often the basis of innovations in the public and private sectors, as they can elaborate new business models, the digitalisation of services, or product customisation. In addition to yielding the economic benefits motivating data-based innovation, these novelties have (un)desirable social implications. For example, a desirable implication relates to the transparency node, an ideal pursued by many legislations encouraging public institutions to share their data openly [41]. Transparency and participatory governance, allowing citizens and other stakeholders to monitor and take part in government

initiatives, is a frequent driver. But so is the social and commercial value of data. Indeed, governments also decide to open their data to capitalise on its commercial value, allowing stakeholders to reuse and innovate upon it. Our model does recognise the potential of open data and transparency to fuel innovation while also considering the possible undesirable implications, such as misuse of open data, which can lead to the creation of dirty data.

Other undesirable implications relate to the shift to the digital delivery of products and services, which has been documented to feed the digital divide, the gap between those with access to technology and digitalised products and services and those without [42]. Such a trend reinforces existing imbalances and isolation dynamics to the detriment of certain shares of the population, which are generally worse off than the others, as being poorer, part of minorities, or older, for example [42]. New analytical techniques have come with many challenges in recent years, magnifying exiting problems or introducing new ones. Automated algorithmic decision-making, today widely used in many different contexts, has been recognised to amplify further existing systematic issues [43,44], while having the novel characteristics of being "unregulated and often hidden and invisible" ([45], p. 394), which makes it more difficult to obtain accountability for the problems it generates. The perverse effects of algorithm use are often attributed to the data on which the models were trained. Indeed, as algorithms use their training data to make decisions, they are likely to reproduce patterns encountered in the training dataset, which may be dirty, outdated, or reflect biases from the past [46,47]. We argue that if dirty data feeds algorithmic harm, the opposite is true, as harmful algorithmic predictions will be stored in databases and feed the overall volume of data. The growing exposure to data and information can result in an overload having adverse effects on mental and physical health and productivity [48]. Moreover, the increasing quantity of dirty data also introduces risks of losing the ability to tell correct information from fake news and can hence lead to disinformation [48]. Lastly, the model considers that data-driven innovations can have different types of effects on the consumption of energy. Indeed, data can be used to innovate in many ways, and while some of these will enable efficiency gains, others will further increase energy needs.

4.4 Summing up: What is data pollution?

While the nodes within the model interact to constitute the feedback loops, we need to recognise the interactions among the feedback loops themselves and their respective sections. Section A represents the mechanisms propelling the rapid expansion of the data economy, which directly influences the requisites for data storage and infrastructure, as addressed in section B. As the volume of data continues to increase, it results in a surge in the demand for data storage and infrastructure. This growth in data storage typically implies higher energy consumption, leading to adverse environmental consequences. In this context, the social implications elucidated in section C gain prominence. The social implications and its associated innovations are inextricably linked to the growth in data volume. As data volume increases, it serves to magnify the social challenges and disparities mentioned in section C, thereby reinforcing the interactive nodes at play within the conceptual model. We hence come to the following synthetic definition:

> Data pollution is the set of harms generated by economic activities related to data collection, storage and use, which transversally impacts individuals and their living environments.

5 Current policy responses to data pollution

Overall, the complexity of data pollution arises from the interactions between its components, represented as nodes in our model, which are not easily separable given their circularity formed through feedback loops. This is especially true given that an effect can feed back into its cause. Because of this complexity, we believe it is necessary to tackle it comprehensively. However, to do so, there is a need to understand what is currently in place in the EU to address the complex mechanisms generating data pollution.

Firstly, and considering the rapid expansion of the data-driven economy, the EU has adopted a Data Strategy in 2020 aiming to foster data economy by creating a single data market, enhancing Europe's competitiveness and data sovereignty [49,50]. The Data Governance Act and the Data Act serve as a legal framework to enable the practical implementation of the EU Data Strategy [51,52]. The Data Governance Act aims to reinforce the single market for data notably by setting the conditions for common European data spaces aiming to facilitate the exchange of public sector and business data across the EU to stimulate the growth of novel data-driven products and services [5,6]. Indeed, as the volume of data generated from digital devices and services continues to grow, there was a need for a legal framework aiming to harness the potential of data to the benefit of the EU economy and society, and to avoid reliance on third countries, particularly in the advancement of IoT or AI systems [53]. In this aspect, the Data Act complements the Data Governance Act by setting rules for using data generated by IoT to boost the EU's data economy [54]. The Data Act also fits the idea of growing the EU data-driven economy by creating avenues and eliminating obstacles for data reuse [55]. Complementarily for the public sector and publicly funded data, the Open Data Directive completes the desire of the EU to foster data sharing and data-driven innovation across all sectors of the economy by promoting the availability and reusability of public sector data in the EU [56]. The Open Data Directive is a central instrument for the realisation of the EU data economy by underlying the willingness of the EU to capitalise on public sector data to feed the data economy [57]. By doing so, it also gives the right to individuals to access information retained by public authorities, facilitating government transparency and accountability, nurturing public trust, and, in turn, expanding public engagement [58]. While these documents do not explicitly address data pollution, they still contain elements and objectives that may indirectly contribute to mitigating the dark sides of the data-driven economy and the data pollution it generates. For example, they encourage responsible data sharing through European data spaces, which may reduce data fragmentation and inefficiency, contributing to data pollution. They also encourage responsible data management practices such as the publication of high-quality data, which indirectly reduce the risk of data pollution arising from the dissemination of dirty data and reduce the risks of misuse, such as the creation of bias or inaccurate analysis contributing to misinformation. Through their broader goals of promoting responsible data management practices, these documents can indirectly contribute to reducing data pollution. There is however no specific provision addressing data pollution as a standalone issue.

Secondly, as the volume of data increases, the demand for infrastructure and user hardware also grows in proportion to the data volume. However, data storage activities generate data pollution by having ecological consequences. The EU only addresses indirectly data pollution generated by data storage only through broader environmental and sustainable efforts. For example, the Circular Economy Action Plan encourages responsible product design and management, including electronic and IT equipment [59]. Other examples are the Regulation on Ecodesign Requirements for Servers and Data Storage Products, which aims to limit the environmental impacts of servers and data storage products by setting rules on energy efficiency [60] or the revised Energy Efficiency Directive, which tackles the heating and cooling of data centres by introducing an obligation for the monitoring of the energy performance of data centres to ensure a fully decarbonised heating and cooling supply by 2050 [61]. These examples show how environmental aspects of data pollution related to data storage are typically addressed through broader regulations setting energy efficiency targets and emissions reduction goals or tackling distinct problems such as responsible data management (e.g., GDPR, Digital Service Act (DSA), Digital Market Act (DMA)) or the recycling of electronic equipment used for data storage through the Waste Electrical and Electronic Equipment Directive [62-65].

Thirdly, the quest to foster a data-driven economy has created social implications by generating an environment in which some data giants have unique control over some data, which is no longer offset by the control of other actors [50]. As technology has become increasingly complex and invasive, especially given that various businesses exploit data through algorithmic decision-making, it is increasingly complicated for consumers and users to maintain control over their data [50]. This led to the emergence of a darker narrative around the data-driven economy, especially as individuals encounter daily vast amounts of information across various devices and media, which can overwhelm them and thus jeopardise their ability and motivation to scrutinise essential details for informed decisions and instead chose

defaults' options which are presented to them as recommendations [50]. This leads individuals to often provide consent without considering the consequences, especially when faced with consent requests. Considering that data are nonrival and the potential benefit they hold for the economy, personal data are widely shared, reused and hence risk being misused. To address these concerns, the EU introduced the GDPR in 2018, aiming to regulate the processing of personal data within the EU, granting consumers enhanced protection and control over their data [63]. The GDPR sets standards for data privacy by covering any organisation that collects or processes EU citizens' data independently of the organisation's location [66]. The GDPR primarily focuses on protecting individuals' data, ensuring their privacy rights are respected and providing mechanisms for individuals to control how their data are collected, processed, and used [66]. Other regulations have entered into force to strengthen online rights and regulate digital services, such as the DSA and the DMA [64,65]. The DSA aims to regulate digital platforms to protect individuals and their fundamental rights online while fostering innovation, growth, and competitiveness. The DMA is another legislative framework targeting digital platforms to ensure fair competition. While these legislative frameworks primarily focus on protecting rights and regulating digital services rather than addressing data pollution as a separate issue, they include some provisions that may indirectly help address some aspects of data pollution. For example, the GDPR promotes the right to erasure, which ensures that personal data can be erased under some circumstances, such as if the personal data are no longer necessary for the purposes for which they were collected. Another illustrative example is the principle of data minimisation, which encourages organisations to collect and process only the data that is directly relevant and necessary to accomplish a specified purpose. Hence, by intending to limit excessive data collection, the GDPR may reduce the data pollution generated by accumulating unnecessary data. As for the DSA, regulating digital platforms, notably through content regulation and platform accountability, it indirectly addresses data pollution through provisions aiming at reducing harmful or misleading data, leading to the spread of illegal content and disinformation. Finally, while the primary objective of the DMA is to address competition issues in the digital market, it also indirectly addresses data pollution by fostering responsible data management and ensuring fair access to data.

6 Conclusion

Data pollution is a term that has been used since the 1980s in different domains and contexts and has hence been given different meanings over time. As data pollution has always been considered within some specific contexts of interest of the authors who wrote about it, but never in its globality, it has only enabled the conceptualisation of partial solutions to the set of harms entailed in data pollution. However, while being different, the meanings attributed to data pollution have a similarity: they describe harms produced as byproducts of the data economy. With this article, we propose a conceptualisation of data pollution, positioning it within the context from which it emerges and that perpetuates, it and contending that all the different harms produced by the data activities not only originate from interdependent mechanisms but also result from and feed one another. As a result of this conceptualisation, we define data pollution as the set of harms originating from any data activity. We believe such a holistic approach is necessary for policymakers to conceive ways to address data pollution effectively.

We argue that the current EU regulation primarily focuses on specific aspects of data pollution while ignoring other large strands of the phenomenon and is not fit to grant adequate protection against many of the harms of data pollution. Although concerns over data subjects' privacy and considerations over the competitiveness of infrastructures and European companies are addressed by the law, ecological harms caused by data activities and many social adverse effects, including data exclusion and information overload, are not. Indeed, while the various law texts each address specific aspects of data pollution, their provisions are not directly related and do not account for the interrelated nature of the harms stemming from data pollution. To ensure that the continuous digitalisation of our activities allows the transition to a more ecological and socially fair society, we it is essential to understand data pollution comprehensively and address it as such, with a coordinated regulatory approach. We hope that the

concept of data pollution as presented in this article can serve as the basis for future research investigating the possibilities to create regulation addressing data pollution comprehensively.

Acknowledgments. This study was funded by the Swiss National Science Foundation (grant number 212637).
Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Tan, K.H., Ji, G., Lim, C.P., Tseng, M.L.: Using big data to make better decisions in the digital economy, International Journal of Production Research 55 (17), 4998-5000 (2017)
2. Calzada, I., Almirall, E.: Data ecosystems for protecting European citizens' digital rights, Transforming Government: People, Process and Policy 14 (2), 133-147 (2020)
3. Curry, E., et al.: The European big data value ecosystems, in: E. Curry, A. Metzger, S. Zillner, J.C. Pazzaglia, A.G. Robles (Eds.), The Elements of Big Data Value, pp. 3-20, (2021)
4. The Economist Group: The future of Europe's data economy, https://impact.economist.com/perspectives/sites/default/files/ei233_msft_futuredata_report_-_v7.pdf, last accessed 2023/11/08 (2022)
5. European Commission: Building a data economy - Brochure, https://digital-strategy.ec.europa.eu/en/library/building-data-economy-brochure, last accessed 2023/11/08 (2019)
6. Common European data spaces, https://dataspaces.info/common-european-data-spaces/#page-content, last accessed 2023/11/08
7. Andersen, K.N., et al.: Fads and facts of e-government: A review of impacts of e-government (2003–2009), International Journal of Public Administration 33 (11), 564-579 (2010)
8. Baku, A.A.: Digitalisation and new public management in Africa, New Public Management in Africa: Contemporary Issues, pp. 299-316, (2022)
9. Global Open Data Index, http://index.okfn.org/place.html, last accessed 2024/03/04
10. ThoughtLab: Building a Future-Ready City, https://thoughtlabgroup.com/wp-content/uploads/2022/11/Building-a-Future-Ready-City-ebook_Final-November-2022.pdf, last accessed 2024/03/04 (2022)
11. Ben-Shahar, O.: Data Pollution, Journal of Legal Analysis 11, 104-159 (2019)
12. Hasselbalch, G.: Data Pollution & Power: White paper for a global sustainable agenda on AI, https://gryhasselbalch.com/books/data-pollution-power-a-white-paper-for-a-global-sustainable-development-agenda-on-ai/, last accessed 2023/11/08 (2022)
13. The Shift Project: Lean ICT - Towards Digital Sobriety, https://theshiftproject.org/wp-content/uploads/2019/03/Lean-ICT-Re%ADport_The-Shift-Project_2019.pdf, (2019)
14. Amoroso, D.L., Mcfadden, F., White, K.B.: Disturbing realities concerning data policies in organizations, Information Resources Management Journal (IRMJ) 3 (2), 18-28 (1990)
15. Zimmerli, W.C.: Who is to blame for data pollution? On individual moral responsibility with information technology, Philosophy and Technology II: Information Technology and Computers in Theory and Practice, Springer, pp. 291-305, (1986)
16. Cao, Y., et al., Efficient repair of machine learning systems via causal unlearning, In: J. Kim, et al. (eds.), Asia conference on computer and communications security, pp. 735-747. Association for Computing Machinery, Incheon, Republic of Korea (2018)
17. Esfahani, A., Mantas, G., Rodriguez, J., Nascimento, A., Neves, J.C., A null space-based MAC scheme against pollution attacks to random linear network coding, In: O. Swantee, J. Wang (eds.), International Conference on Communication Workshop, pp. 1521-1526. IEEE, London, UK (2015)
18. Chang, J., et al.: Application of deep machine learning for the radiographic diagnosis of periodontitis, Clinical Oral Investigations 26 (11), 6629-6637 (2022)
19. De Nadai, A.S., Hu, Y., Thompson, W.K.: Data pollution in neuropsychiatry—an under-recognized but critical barrier to research progress, JAMA psychiatry 79 (2), 97-98 (2022)
20. Lin, H., Ai, Y., Ling, Z., A light CNN with split batch normalization for spoofed speech detection using data augmentation, In: N. Theera-Umpon, et al. (eds.), Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, pp. 1684-1689. IEEE, Chiang Mai, Thailand (2022)

21. Dong, J., Curtmola, R., Nita-Rotaru, C., Yau, D.K.Y.: Pollution attacks and defenses in wireless interflow network coding systems, Transactions on Dependable and Secure Computing 9 (5), 741-755 (2012)
22. Castelluccia, C., Kaafar, M.A., Owner-centric networking (onc): Toward a data pollution-free internet, In: K. Yoshida, M. Chang (eds.), Ninth Annual International Symposium on Applications and the Internet, pp. 169-172. IEEE, Bellevue, USA (2009)
23. Shenk, D.: Data smog: Surviving the information glut, Harper, San Francisco (1998)
24. Gouws, R.H., Tarp, S.: Information overload and data overload in lexicography, International Journal of Lexicography 30 (4), 389-415 (2017)
25. Bawden, D., Holtham, C., Courtney, N., Perspectives on information overload, Aslib Proceedings, vol.51, pp. 249-255. (1999)
26. "Climate crisis: The unsustainble use of online video": Our new report on the environmental impact of ICT, https://theshiftproject.org/en/article/unsustainable-use-online-video/, last accessed 2023/10/24
27. Data centres and data transmission networks, https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks, last accessed 2023/10/24
28. Bietti, E., Vatanparast, R.: Data Waste, Harvard International Law Journal Frontiers 61, 1-11 (2020)
29. François, C.: Systemics and cybernetics in a historical perspective, Systems Research and Behavioral Science: The Official Journal of the International Federation of Systems Research 16 (3), 203-219 (1999)
30. Heylighen, F., Joslyn, C.: Cybernetics and second-order cybernetics, Encyclopedia of physical science & technology, pp. 155-170, (2001)
31. Mitchell, M.: Complex systems: Network thinking, Artificial intelligence 170 (18), 1194-1212 (2006)
32. Glanville, R.: Second order cybernetics, Systems Science and Cybernetics 3, 59-85 (2002)
33. Scott, B.: Second-order cybernetics: An historical introduction, Kybernetes 33 (9/10), 1265-1378 (2004)
34. Rose, K., Eldridge, S., Chapin, L.: The internet of things: An overview, https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-IoT-Overview-20151221-en.pdf, last accessed 2023/11/08 (2015)
35. Windows phone was a glorious failure: Looking back on the bumpy road taken by Microsoft's most ambitious mobile OS, https://www.theverge.com/2017/10/10/16452162/windows-phone-history-glorious-failure, last accessed 2023/11/05
36. Rietveld, J., Schilling, M.A.: Platform competition: A systematic and interdisciplinary review of the literature, Journal of Management 47 (6), 1528-1563 (2021)
37. Katz, M.L.: Multisided platforms, big data, and a little antitrust policy, Review of Industrial Organization 54 (4), 695-716 (2019)
38. Papadopoulou, E., Stobart, A., Taylor, N.K., Williams, M.H., Enabling data subjects to remain data owners, In: G. Jezic, R. Howlett, L. Jain (eds.), Agent and Multi-Agent Systems: Technologies and Applications: 9th KES International Conference, pp. 239-248. Springer International Publishing, Sorrento, Italy (2015)
39. European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: "Towards a common European data space", COM(2018) 232 final, Brussels, Belgium, 2018, pp. 1-15.
40. Google: Environment Report, https://www.gstatic.com/gumdrop/sustainability/google-2023-environmental-report.pdf, last accessed 2023/10/26 (2023)
41. Attard, J., Orlandi, S., Scerri, S., Auer, S.: A systematic review of open government data initiatives, Government Information Quarterly 32 (4), 399-418 (2015)
42. Cullen, R.: Addressing the digital divide, Online Information Review 25 (5), 311-320 (2001)
43. Donovan, J., Caplan, R., Matthews, J., Hanson, L.: Algorithmic accountability: A primer, https://datasociety.net/wp-content/uploads/2019/09/DandS_Algorithmic_Accountability.pdf, last accessed 2023/11/08 (2018)
44. Sepúlveda Carmona, M., Secretary-General, U., Poverty, U.S.R.o.t.Q.o.H.R.a.E.: Extreme poverty and human rights: Note by the Secretary-General, https://digitallibrary.un.org/record/1648309?ln=en, last accessed 2023/11/08 (2011)
45. Marjanovic, O., Cecez-Kecmanovic, D., Vidgen, R.: Algorithmic pollution: Making the invisible visible, Journal of Information Technology 36 (4), 391-408 (2021)
46. The police are using computer algorithms to tell if you're a threat, https://time.com/4966125/police-departments-algorithms-chicago/, last accessed 2023/11/10
47. Rosenblat, A., Wikelius, K., Boyd, D., Gangadharan, S.P., Yu, C., Data & civil rights: Criminal justice primer, In: D. Boyd, S. Peña Gangadharan, C. Yu (eds.), Data & Civil Rights Conference, Washington, DC (2014)
48. Bawden, D., Robinson, L.: Information overload: An overview, Oxford Encyclopedia of Political Decision Making, Oxford University Press, Oxford, (2020)

49. European data strategy, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en#data-governance, last accessed 2023/10/26
50. Van Ooijen, I., Vrabec, H.U.: Does the GDPR enhance consumers' control over personal data? An analysis from a behavioural perspective, Journal of consumer policy 42, 91-107 (2019)
51. European Commission, Proposal for a Regulation of the European Parliament and of the Council on harmonised rules on fair access to and use of data (Data Act), 2022.
52. Official Journal of the European Union, Regulation (EU) 2022/868 if the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 018/1724 (Data Governance Act), in: Official Journal of the European Union (Ed.) 2022, pp. 1-44.
53. Baloup, J., et al.: White paper on the data governance act, https:// papers.ssrn.com/sol3/papers.cfm?abstract_id=3872703, last accessed 2023/11/08 (2021)
54. European Commission, Data Act: Commission welcomes political agreement on rules for a fair and innovative data economy, European Commission, Brussels, 2023.
55. The European Data Act, https://www.eu-data-act.com/, last accessed 2023/11/07
56. van Eechoud, M.: A serpent eating its tail: The database directive meets the open data directive, IIC-International Review of Intellectual Property and Competition Law 52 (4), 375-378 (2021)
57. Official Journal of the European Union, Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information, in: Official Journal of the European Union (Ed.) 2019, pp. 1-28.
58. Broomfield, H.: Where is open data in the Open Data Directive?, Information Polity 28 (2), 175-188 (2023)
59. European Commission, Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: A new Circular Economy Action Plan - For a cleaner and more competitive Europe, Brussels, 2020.
60. Official Journal of the European Union, Commission Regulation (EU) 2019/424 of 15 March 2019 laying down ecodesign requirements for servers and data storage products pursuant to Directive 2009/125/EC of the European Parliament and of the Council and amending Commission Regulation (EU) No 617/2013, in: Official Journal of the European Union (Ed.) 2019, pp. 1-21.
61. Energy efficiency directive, https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficiency-targets-directive-and-rules/energy-efficiency-directive_en#heating-and-cooling-and-data-centres, last accessed 2023/11/07
62. Official Journal of the European Union, Directive 2012/19/EU of the European Parliament and of the Council of 4 July 2012 on waste electrical and electronic equipment (WEEE), in: Official Journal of the European Union (Ed.) 2012, pp. 1-34.
63. Official Journal of the European Union, Regulation (EU) 2016/679 of the European Parliament and the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016, pp. 1-88.
64. Official Journal of the European Union, Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act), in: Official Journal of the European Union (Ed.) 2022, pp. 1-102.
65. Official Journal of the European Union, Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act), in: Official Journal of the European Union (Ed.) 2022, pp. 1-66.
66. Zaeem, R.N., Barber, K.S.: The effect of the GDPR on privacy policies: Recent progress and future promise, ACM Transactions on Management Information Systems (TMIS) 12 (1), 1-20 (2020)