Original Research Article

BIG DATA & SOCIETY

Big Data & Society
July–December: 1–10

DOI: 10.1177/20539517221139785
journals.sagepub.com/home/bds
SAGE

Politics of data reuse in machine learning systems: Theorizing reuse entanglements

Nanna Bonde Thylstrup[1], Kristian Bondo Hansen[1], Mikkel Flyverbom[1] and Louise Amoore[2]

Abstract

Policy discussions and corporate strategies on machine learning are increasingly championing data reuse as a key element in digital transformations. These aspirations are often coupled with a focus on responsibility, ethics and transparency, as well as emergent forms of regulation that seek to set demands for corporate conduct and the protection of civic rights. And the Protective measures include methods of traceability and assessments of 'good' and 'bad' datasets and algorithms that are considered to be traceable, stable and contained. However, these ways of thinking about both technology and ethics obscure a fundamental issue, namely that machine learning systems entangle data, algorithms and more-than-human environments in ways that challenge a well-defined separation. This article investigates the fundamental fallacy of most data reuse strategies as well as their regulation and mitigation strategies that data can somehow be followed, contained and controlled in machine learning processes. Instead, the article argues that we need to understand the reuse of data as an inherently entangled phenomenon. To examine this tension between the discursive regimes and the realities of data reuse, we advance the notion of reuse entanglements as an analytical lens. The main contribution of the article is the conceptualization of reuse that places entanglements at its core and the articulation of its relevance using empirical illustrations. This is important, we argue, for our understanding of the nature of data and algorithms, for the practical uses of data and algorithms and our attitudes regarding ethics, responsibility and regulation.

Keywords

Data reuse, machine learning, ethics, entanglements, datasets, algorithms

Introduction

Policy discussions and corporate strategies on machine learning are increasingly championing data reuse as a key element in digital strategies. These aspirations are often coupled with a focus on responsibility, ethics and transparency, as well as emergent forms of regulation that seek to set demands for corporate conduct and the protection of civic rights, such as the right to privacy and the right of erasure of 'data subjects'. In most contexts, the drive for more data reuse and the demand for more accountability and protective measures seem to go hand in hand. We can apparently have both if we ensure that ethics and responsibility measures are appropriately considered, data can travel and be reused across sectors and territories, and algorithms from one domain can be installed elsewhere unproblematically. Protective measures include methods of traceability and assessments of 'good' and 'bad' datasets and algorithms that are considered to be traceable, stable and contained.

However, these ways of thinking about both technology and ethics overlook a fundamental issue that this article seeks to address, namely that machine learning systems entangle data, algorithms and more-than-human environments in ways that challenge a well-defined separation, that is, the deep entanglement between data, algorithms and sociotechnical infrastructures render the imaginaries of transparency, traceability and clean boundaries between data and algorithms in machine learning impossible.

The understanding of reuse that we investigate underpins many facets of contemporary digital transformations among practitioners in different industries and policy circles. Data reuse is often encouraged by open source and science

[1]Department of Management, Society and Communication, Copenhagen Business School, Frederiksberg, Denmark
[2]Department of Geography, Durham University, Durham, UK

Corresponding author:
Nanna Bonde Thylstrup, Department of Management, Society and Communication, Copenhagen Business School, Frederiksberg, Denmark.
Email: nbt.msc@cbs.dk

ideologies that promise to kinetically 'unleash' (Mayer-Schönberger and Cukier, 2013) and 'unlock' (Verhulst, 2020) the value of otherwise 'hidden' and 'dormant' data. Corporations have made considerable investments in establishing policies, infrastructures and economies that incentivize and undergird data reuse. These investments rely on data reuse strategies that frame data as discrete, interchangeable and governable commodities that must be moved up and down different supply chains to create new value (Alaimo et al., 2020; Sands, 2018). Data-driven corporations like Palantir (Alden, 2018) and consulting companies in the field of AI such as McKinsey (Fountaine et al., 2021) promise to 'unmoor' data from the 'information silos' that currently constrain them. Similarly, Google repurposes data logs ('digital exhaust') into predictive patterns (Zuboff, 2019) and IBM reuses algorithms built for seismic detection to develop predictive policies (Barnes and Wilson, 2014; Hälterlein, 2021).

At the policy level, data reuse also forms a cornerstone in policy and innovation discourses on AI and machine learning. The European Commission recently presented its formal draft of the EU Data Act in February 2022. A core component of this act is to facilitate easier data sharing and use/reuse by setting standards at the EU level. Reuse is also an important element in the EU's open data directive (Directive (EU) 2019/1024, 2019). It subtends almost all discussions on privacy (Custers and Uršič, 2016a) and scientific knowledge infrastructures (Pasquetto et al., 2017; Van De Sandt et al., 2019). It has become a core argument for the advancement of data-driven economies (Custers and Bachlechner, 2017) and a new weapon in the arsenal of public health technologies (Madianou, 2020). Data reuse is even inscribed as a fundamental scientific principle for digital data management in the sciences (Wilkinson et al., 2016). The justification for these varied mobilizations of reuse is typically utilitarian, framing the non-use of data as a waste of resources. In this logic, if data exists, it should also be made available for further value extraction across both the public and private sectors (Zuboff, 2019). Data in this sense also intersects with corporate and governmental desires for growth and innovation through AI (Amoore, 2013).

The aspirations for data reuse and algorithmic solutions – couched in responsibility, privacy and transparency – overlook the issues identified in this article, namely that data, algorithms and technologies are fundamentally entangled and cannot be decoupled. We consider this perspective to be important because it expands the political and ethical register of data reuse and delineates why and how current reuse strategies will have to contend with this expanded register.

The importance of data and model reuse in data-driven societies has also resulted in new controversies. Ethical debates have emerged around the production and circulation of training data (Andrew et al., 2020; Birhane et al., 2021; Harvey and LaPlace, 2021; Thylstrup, 2022). Numerous scholars have called for new guidelines on the prevention and reuse of contested data in research projects (Van Noorden, 2020). Current mitigation strategies for the reuse of contested data rely on the strategies of deletion, containment and transparency. However, it has been highlighted in several studies that these responses are associated with their own set of challenges because iterations of datasets are still present even after the original set is removed (Peng, 2020). Moreover, the methods of tracing and accounting for harmful datasets, for instance through dataset auditing practices, comes with its own methodological challenges and ethical conundrums (Raji et al. 2020; Keyes and Austin 2022). The challenges of hindering the reuse of contested data have also engendered new strategies to protect corporations against data as a 'toxic asset' (Schneier, 2016), and new discourses that frame sensitive data as potentially 'toxic waste' that can 'overspill' (Schwarzkopf, 2020). Nevertheless, researchers in the field of machine learning acknowledge the mismatch between the imaginary of data as discrete entities, as expressed in both reuse strategies and data regulation, and the reality of how data is utilized in machine learning processes. As noted in a leaked internal document written by Facebook privacy engineers about the company's challenges in dealing with user data and privacy regulations:

> "We've built systems with open borders. The result of these open systems and open culture is well described with an analogy: Imagine you hold a bottle of ink in your hand. This bottle of ink is a mixture of all kinds of user data (3PD, 1PD, SCD, Europe, etc.) You pour that ink into a lake of water (our open data systems; our open culture) … and it flows … everywhere. How do you put that ink back in the bottle? How do you organize it again, such that it only flows to the allowed places in the lake?"

This article seriously considers the concerns of the Facebook engineers, namely the fundamental fallacy of most data reuse strategies as well as their regulation and mitigation strategies that data can somehow be followed, contained and controlled in machine learning processes. Instead, the article argues that we need to understand the reuse of data as an inherently entangled phenomenon. To examine this tension between the discursive regime and the reality of data reuse, we advance the notion of reuse entanglements. We use this notion as a conceptual lens, and introduce the issue of data reuse into a conversation on the social theories of entanglement.

Analytically, our article argues for the need to shift from the linear framing of data reuse found in corporate and governmental strategies to an emphasis on reuse as an inherently iterative and entangled phenomenon in machine learning regimes. Conceptually, such a shift also creates a

broader space for discussions on data and algorithmic ethics that consider their mutual constitution instead of their categorization as mere problems of input and output. Although dominant AI ethics frameworks rely on the assumptions of the existence of 'ethical data' (Mittelstadt and Floridi, 2016; Zook et al., 2017) and 'ethical algorithms' (Mittelstadt et al., 2016), we instead mobilize the notion of reuse entanglements to emphasize the impossibility of defining such distinct spaces between data and algorithms and 'good' and 'bad' in machine learning regimes. A notable example is the recent partnership between the United Nations' World Food Programme (WFP) and the controversial US software company Palantir, which specializes in the areas of predictive policing and surveillance, among others. Although the WFP celebrated the partnership as tech-for-good, arguing that the relationship was without concern because Palantir would not allow access to the information of beneficiaries, an entanglement perspective suggests that such partnerships should be more critically scrutinized. Even if the data is non-personal, the reuse of data related to the movements of refugees by a defence contractor, for example, still raises ethical and political issues, such as the spill-over effects of Palantir's models (Martin et al., 2022). The same applies to algorithms that may appear to be low risk but may contain traces of deeply contested data. Another example is the increasingly significant role software libraries play in, for instance, the military. A recent report published by National Security Commission on Artificial Intelligence (authored by Eric Schmidt among others), thus describes software libraries such as Pytorch and Tensorflow as important parts of the US military strategy (United States, 2021). It emphasizes the crucial role reuse platforms play in 'transforming research prototypes to production-ready machine learning models'. These platforms thus raise ethical and political questions about reuse that exceed more simple assessment of whether data was ethically harvested or a model ethically constructed.

In the following, we begin by providing an overview of existing discussions on how machine learning practices shape the concepts, practices and politics of data reuse. We then mobilize theories on entanglement to develop an analytical framework of reuse that considers its intra-relational nature. Third, we bring science and technology studies studies on data reuse practices into conversation with Karen Barad's framework of entanglement to discuss empirical examples of the boundary-making practices of data science that configure entangled data phenomena into discursive discrete entities. Finally, we remark on the ethico-politics of reuse in machine learning systems.

The main contribution of this article is the conceptualization of reuse that places entanglements at its core and the articulation of its relevance using empirical illustrations. This is important for our understanding of the nature of data and algorithms, for the practical uses of data and algorithms, and our attitudes regarding ethics, responsibility and regulation.

Mapping debates and discourses on data reuse in machine learning systems

Three clearly discernible, but interrelated, bodies of literature have shed light on how machine learning regimes have transformed the definitions, practices and political contexts of data reuse. Each of these strands of literature indicates the need for a theoretical framework of data reuse that counters the framing of reuse processes as linear and discrete. They each respectively challenge three fundamental assumptions inherent in reuse strategies and discourses that data reuse is a linear process that is stabilized based on the conceptual distinction between use and reuse; that data can be discrete and 'raw'; and, that reuse can be apolitical.

Taxonomy

Recent works on information studies and law have shown that machine learning regimes challenge the definition of reuse. Traditionally, two baseline understandings of data reuse have dominated the field: 'the use of data collected for one purpose to study a new problem' (Zimmerman, 2008: 634) and the 'usage of a dataset by someone other than the originator' (Pasquetto et al., 2017: 3). Recently, however, scholars have called for an updated conceptual framework that seriously considers the dynamic and iterative reality of machine learning. Van de Sandt et al. (2019) argue that existing concepts of data reuse rely on paper-centred, linear research models that are far removed from the machinic processes of machine learning regimes. Therefore, they propose to deconstruct the notion of reuse using the concept of '(re)use', which they employ to address 'any research resource regardless of when it is used, the purpose, the characteristics of the data and its user' (Van De Sandt et al., 2019). We find similar gestures of conceptual de- and reconstruction in the field of tech and law, which is adjacent to, but largely not in conversation with information studies. Custers and Uršič, for example, have developed a legal theoretical framework to tackle the issue of data reuse in the age of machine learning by developing a general taxonomy of data reuse that distinguishes between data recycling, data repurposing, data recontextualization, data sharing and data portability (Custers and Uršič, 2016b). Both examples show how machine learning regimes destabilize the very conceptual ground upon which reuse stands.

Scientific practice

Just as machine learning undermines the established conceptual framework of reuse, it also affects the sociotechnical processes and practices of reuse (Aaen et al., 2021;

Winkler and Berenbon, 2021; Tenopir et al. 2015). As scholars from disciplines such as STS and the philosophy of science have shown, data reuse practices have traditionally been shaped by the epistemic cultures (Leonelli and Tempini, 2020) and settings (Loukissas, 2019) that produced the data in the first place. These epistemic cultures have traditionally not only promoted a common contextual understanding of data within disparate scientific fields, for instance, plant phenomics and health research, but also defined the parameters of reuse. Many branches of data science, however, operate under the assumption that data can be recontextualized and that machine learning attains generalizability and circulation through reuse (Ribes et al., 2019). In this epistemic culture, data scientists seek to 'empty' pre-trained algorithms of their domain-specific content to facilitate a seamless transition to new domains, wherein adjustment and modification can be performed in situ (Ribes et al., 2019). These recent works on reuse practices in the data sciences help us gain a better understanding of the practices and epistemic implications of unmooring data from their context.

Politics

In recent years, more and more scholars have begun to pay attention to how data reuse in machine learning regimes is embedded in racialized and gendered structures, and how these embeddings give rise to new questions related to labour, power and consent (Cifor et al., 2019; Mulvin, 2021; Radin, 2017; Sutherland, 2021). These works challenge the aforementioned decontextualization mechanisms of data science by emphasizing the significance of context (Radin, 2017), and focusing on questions of reuse related to data justice and data violence (Stevens and Keyes, 2021; Thylstrup et al. 2021). Research on global data justice, for instance, has shown how data reuse for AI development often occurs via hidden and opaque practices in extra-legal spaces where data protection is scarce (Martin et al., 2022). Moreover, critical works on training data have demonstrated the deeply contested reuse entanglements that underpin seemingly neutral standards such as those of the NIST for facial recognition technologies (Keyes et al., 2019). The vulnerabilities and risks identified by this body of work highlight the tension between the scientific ideologies of open source and the societal structures of power in the development of machine learning systems, echoing the politics of open access on a broader level, where 'accessibility is synonymous with "open to all" without regard to cultural, social and historical conditions' (Christen, 2018).

These three bodies of work are relevant to understanding why 'data reuse' has become a contested cornerstone in political and corporate reuse strategies, and the role of scientific practice in this realm. We can infer from these diverse strands of research that the machine learning environment fundamentally transforms the conceptual, epistemic and political landscape of data reuse, and that it now hinges on a generative process between data, algorithms and more-than-human environments. This highlights an inherent tension between the sociotechnical imaginaries of reuse that enact the logics of linearity and isolation, and their less well-defined reality. In the next section, we examine these perspectives and highlight 'reuse' not simply as a technical challenge, but also as a fundamentally entangled socio-technical phenomenon. This requires some conceptual articulation that we now consider.

Conceptualizing reuse entanglements in machine learning regimes

This article develops the notion of reuse entanglements to challenge the understanding of data as 'raw' and amenable to 'linearity' in machine learning practices. What is at stake in placing the concept and practice of data reuse within the framework of entanglement? At its most fundamental, this it argues that data as phenomena do not pre-exist in the techno-scientific apparatus in which they appear, since 'any time entities interact they entangle' (Gilder, 2008: 3). This also means rejecting the mainstream misconception of data as separate entities that can be made to relate or interact with one another at will. Karen Barad has expressed this problem in her extended engagement with quantum entanglement, in which she notes that 'phenomena are the ontological inseparability of intra-acting agencies. That is, phenomena are ontological entanglements' (2007: 333). At the core of Barad's argument is a sense of how entities that may appear to be individual – such as data – actually 'emerge through and as part of their entangled intra-relating' (2007: ix).

Barad's work enables us to highlight how data does not pre-exist its use and reuse, but is instead constituted through its entanglements in the broader 'experimental arrangement'. The reuse of data is thus always part of a larger experimental arrangement of algorithms, models and applications (not in the sense of apps but the sense of specific deployments) that determine what is and is not meaningful. In the case of Barad, this means that the 'boundaries and properties' of data are enacted by 'the agential cut determined by the larger experimental arrangement', wherein this experimental arrangement is the 'condition of possibility for particular concepts to be meaningful at the exclusion of others' (2007: 345). In machine learning, such experimental arrangements span a spectrum of more-than-human environments that include computer scientists, data, algorithms, GPUs, human traces, wires and servers and the contexts range from border security to risk scoring of citizens over ad tech and health assessments in insurance.

Significantly, Barad situates the ontology of separation within a representational framework. Representationalism, argues Barad (2007: 137), establishes the concept of

separation as foundational because it 'separates the world into the ontologically disjunct domains of words and things, leaving itself with the dilemma of their linkage such that knowledge is possible.' Barad defines these practices of representation as material-discursive labour practices that 'help constitute and are an integral part of the phenomena being investigated' (Barad, 2007: 232). Crucially, such representation practices do not simply detect differences, for example between self and other, but also produce and reconfigure moments and structures of differences such that self and other appear as separate entities. Countering this framework, Barad argues that we should instead understand phenomena in the world as entanglements that are constituted through, and ontologically inseparable from, the performative practices of representation.

In the following section, we consider Barad's framework of the ontology of separation to home in terms of how reuse practices in machine learning separate data, algorithms and more-than-human worlds into disjunct 'domains of words and things'. Based on empirical examples of a variety of practical examples, we highlight the value of our conceptualization of reuse entanglements as an analytical approach.

Infrastructures and practices of 'thingification' in reuse entanglements

Western scientific paradigms tend to 'thingify' relations into 'things' and 'entities' (Barad, 2003). We accept these conventions as central to data reuse practices and imaginaries, because they reclassify data as isolated and manageable entities (Alaimo and Kallinikos, 2020) that are amenable to the separation of 'object' from 'subject'. This separation is justified by a Neoplatonic (McQuillan 2018) framing of data science as a domain-independent science that can unlock patterns and puzzles across all sectors and scientific fields (Ribes et al., 2019). As David Ribes et al. (2019) assert, the idea that machine learning is an effective form of science that can be applied on a general level is informed by an understanding of data science as either 'emptied' of domain knowledge or as a method that has assembled enough specific domain knowledge. Pre-trained algorithms that are void of domain-specific content in the early layers may transition more smoothly to new domains, where adjustment and modification can take place in situ (Ribes et al., 2019). During the COVID-19 pandemic, for example, collaborations between DeepMind and the UK Government focused on repurposing 'off the shelf' models that could be 'hacked for use in the NHS'. In such cases, the building of a model involves the hacking and reuse of multiple other configurations of data and algorithms.

This concept of machine learning as 'empty' results in the practice of 'prospecting' within data science, involving 'the work of rendering data, knowledge, expertise and practices of worldly domains available or amenable to engagement with data scientific method and epistemology.' (Slota et al., 2020: 1). To achieve the universalizing properties that can accommodate 'seamless' (Slota et al., 2020: 2) data reuse, all domain-specific markers and characteristics must be stripped off. Hence, the choice of method (i.e. data analytics technique) is reduced to the determination of the structural quality of a dataset (images, text, numbers, audio, etc.), and matching it to a suitable analytics technique (e.g. natural language processing in the case of textual data). In effect, it is this performative practice of separation and representation that is inherent in the incessant demand for more data, models, software and reuse practices that is evident in the contemporary push for data reuse (Slota et al., 2020; see also Fourcade and Johns, 2020 for a similar argument). These practices render entanglements irrelevant and invisible. Let us examine some empirical examples to show how these reuse practices of separation and representation become supported, solidified and institutionalized.

First, there are the emergent infrastructures that support the parcelling, modularizing and sharing of data, code and models across disciplinary and institutional boundaries. One of the most significant propellers of the reuse ethos is the open source movement, which has worked to create an ecosystem of data infrastructures and practices that allow for new machine learning technologies continuously 'assembled from existing, reusable' data (Burke, 2019). This ethos has in recent times in particular been enabled by dataset sharing sites such as Figshare (Plantin et al., 2018) and coding and software libraries such as TensorFlow, Github and PyTorch. Platforms that are often guided by the tenets of open source, and that function as collaborative 'community' spaces that allow developers to 'tinker' and experiment with data and models as well as share new datasets and models. These infrastructures have both been credited with democratizing AI, and facilitating the development of technology for non-computer scientists (Coldicutt, 2018). At the same time, scholars have also recently pointed to their embedding in more centralized infrastructures such as Microsoft (Github) (Franco, 2022), TensorFlow (Google (Hoijtink and Planqué-van Hardeveld, 2022) and PyTorch (Facebook). As such, these sites inhabit and actively co-create grey areas of reuse between community, corporations and state that simultaneously espouse open source ideologies and platformization tendencies. While such positions exists at opposite ends of the spectrum, they nevertheless share the sentiment that data and models are entities that can – and should – be modularized, shared and repurposed. Data repositories and software libraries have thus positioned themselves as crucial nodal points within 'the multiple components and actors organizing data sharing' by offering a comprehensive, flexible ecosystem of tools, libraries and community resources that allow them to 'reach out to' and 'link

together' otherwise scattered actors and institutions of machine learning infrastructures (Plantin et al., 2018).

Second, we can examine how these emergent digital infrastructures facilitate new arrangements and meanings of 'reuse' in deep learning techniques and architectures that reify and undermine linear and parcelled imaginaries of reuse. Examples of these include the deep learning cloud services provided by hyperscalers, such as Amazon Web Services (AWS) and Microsoft Azure or the deep learning processes that occur in tech platforms such as Meta (Narayan, 2022). Much has been written on the privacy implications of the data sharing that occur within these architectures. We, however, are more interested in the new dynamics of reuse that they facilitate and the implications of these dynamics on our understanding of the ethics and politics of reuse in an expanded sense. As Amoore (2020) asserts, deep learning should be understood as architectures that incessantly and recursively enable data and algorithms to meet and entangle in particular ways. The privacy engineer at Facebook (now Meta) cited earlier reflects the implications of this architecture in his description of the difficulties associated with withdrawing, isolating and controlling data, when he notes: 'You pour that ink into a lake of water (our open data systems; our open culture) … and it flows … everywhere. How do you put that ink back in the bottle? How do you organize it again, such that it only flows to the allowed places in the lake?' Thus, these entanglements mean that not only is it impossible to isolate data and trace it back along its pathways, but each exposure also changes the models they interact with which is exactly what is implied by the conceptual framework of entanglement.

Foregrounding the entangled nature of data reuse disavows the idea that humans, data and algorithms can be discerned as separate individual agencies that precede interaction. Instead, it assumes their existence as entanglements in which humans, data and algorithms become mutually constitutive (Barad, 2007). In his study on the use of alternative data sources in the credit scoring practices of banks, Aitken (2017) highlighted this issue of iteration in the machinic processing of big data that is not intuitively relatable to the problem at hand (i.e. credit scoring). For instance, Aitken writes the following in relation to banks' reuse and repurposing of non-credit-related data:

> "Alternative credit scores are far removed from any form of human sight and are, rather, reliant on technologies and mediations (data ingestion mechanisms, algorithmic design, predictive modelling and pattern recognition, machine learning of all types) which operate in spaces which are quite distant from human senses. The trace of these processes that are perceptible to human vision—a final calculative score—is only made visible, can only be seen at all, after opaque processes of aggregation and translation." (Aitken, 2017: 287–288).

Aitken highlights the iterative dynamics of machine learning entanglements, and how its repetitive moments are less marked by fixed sets of processes (despite the assembly line-like connotation invoked by discourses on data supply chains), and more by the continual and opaque reworkings by both human and non-human agencies (on alternative data reuse in finance, see Hansen and Borch, 2022). These reworkings produce ongoing materializations of machine learning entanglements in an open (but not arbitrary) temporal process, wherein the reuse entanglements of machine learning materialize in intra-action with other material-discursive apparatuses (Barad, 2007).

Combined, the three examples outlined above show how contemporary data science discursively reinforces, but materially undermines the idea that reuse entanglements are entities that can be spatially traced from origin to endpoint and – if found to be problematic – deleted at will. Indeed, it is all but impossible to define what is novel about contemporary machine learning without some notion of a reuse of data that exceeds the knowledge that is present in linear or sequential rules. Consider, for example, the common computer science assertion that historical 'hard coded' or 'handcrafted' rules-based systems 'faced difficulties because people struggle to devise formal rules with enough complexity to accurately describe the world' (Chollet, 2018: 2). The precise computational and social problem that machine learning addresses, then, gives rise to its definition in terms of "systems that have the ability to acquire their own knowledge [learning] by extracting patterns from data". To define contemporary machine learning in this way – and notwithstanding the multiple and competing forms of machine learning architectures – as a set of inductive computational practices in which data affords knowledge beyond deductive rules, does undermine a linear notion of data reuse. Put simply, machine learning definitively *requires the entanglement* of data with model in order to function (and to iteratively update its functions).

Thus, although reuse strategies display an understanding of machine learning elements such as data, code and models as discrete and 'closed boxes' that can be managed, coalesced and held separately, reuse practices in machine learning show that these elements are mutually constitutive in recursively unfolding ways. Amoore describes how algorithmic systems 'modify themselves in and through their recursive relations to input data' (2020) in such a way that '[l]ittle pieces of past patterns enter a training dataset and teach the algorithm new things … on and on iteratively, recursively making future worlds'. For example, the use of machine learning in biomedical imaging and radiology implies not merely the training of models on the data of past images, but rather, the generation of new images of potential objects of concern to extract features across multiple data sources that are not strictly present in any single specific original image, nor in hard-coded rules. Though there are many different machine learning architectures involved in a problem such as

biomedical image classification, in all cases the model's architecture also requires that it reuses the data that it does itself generate. For example, in the now ubiquitous backpropagation algorithm, the model repeatedly modifies the weights on the connections so as to minimise the difference between actual output vector and desired output (Goodfellow et al., 2016: 202). In short, machine learning could never simply reuse data in its input layers, for it must also *generate data* for reuse via each pass through the layers – this data itself intimately entangled with outputs. This implies that machine learning does not simply reuse individual and intact past data, but rather, decomposes and recomposes data into non-linear and iterative forms that far exceed the 'import' of data to be reused as input.

The form of iterative learning outlined here shapes the conditions of machine learning entanglements, involving continuous cycles of learning that repeat and modify without an end purpose in mind, beyond the experimental process. Moreover, as the next section points out, it also engenders a new politics of reuse which must be understood not in linear and isolated terms, but rather as recursive entanglements between data, models and more-than-human systems.

The politics of reuse entanglements in machine learning systems: Towards a recursive understanding of reuse

How might we understand reuse entanglements in not only technical, but also political terms? As exemplified by feminist perspectives, the very act of designating something or someone as amenable for use and reuse is a gesture imbued with politics. In her work, 'What's the use?', Ahmed (2019) examines the politics of these 'use' gestures by exploring how using, not using, or being put to use shapes our encounters with the world. Crucially, Ahmed explores the rhetoric of 'function' and its role in colonial and racialized capitalism, identifying how markets have historically divided and organized objects, and subjects and communities into useable or non-useable discrete entities that are amenable to value extraction and biopolitical control. Although Ahmed's work does not directly address digital aspects, her exploration of the genealogies and technologies of use offers a crucial expansion of the politics of use beyond the benevolent optimism of open science. She reminds us that utilitarian justifications of data reuse in industrial and policy discourses often also reproduce the binary idea of usefulness/uselessness, and shows how these ideas conceal deeper mechanisms of potential individual and institutional violence. An example is the utilitarian imperative of not letting 'data go to waste', and the implied assumption of data capitalism that 'data exhaust' is an entity that can and should be freely extracted and exploited for financial gains. Such perspectives can help us to appropriate data reuse strategies by echoing older social philosophies of utility, which were integral to the justification and naturalization of the social orders of colonialism and industrial capitalism for capitalist extraction via labour exploitation and colonial subjectification. Ahmed's perspectives thus resonate with contemporary works on data justice that emphasize gendered, racializing and colonial dimensions of contemporary machine learning systems (Cifor et al., 2019; Hoffmann, 2020; Thatcher et al., 2016). Moreover, they facilitate an understanding of how discussions on the appropriate use, reuse and maintenance of data consider 'how data comes into being' (Radin, 2017) and how 'new types of technology' are imbued with 'old social harms' (Sutherland, 2019). It also sheds light on the politics of framing data as a wasteland that can be cultivated into new property via algorithmic uses (Thylstrup, 2019).

At a deeper level, we also need to assess the political gestures of utilitarianism within a broader set of epistemic and political transformations as states and societies begin to understand themselves and their problems using deep neural network algorithms (Amoore, 2022). Political debates on reuse have focused predominantly on how data is reused for new purposes and in new policy domains that emphasize the transparency of the purpose of data collection, storage, or processing. In the UK, for example, the government is attempting to make all general practice health records available for reuse by researchers and commercial organizations (Kamlana, 2021). However, although primary public and scholarly concerns related to the widespread reuse of health data have been expressed in terms of the repurposing of data for new and unspecified uses, the value to the government and commercial interests is also algorithmic. Access to the world's largest and most structured health dataset is also a route to building and refining machine learning models that cluster and target the patterns of a population. For this reason, it is increasingly important to understand how the reuse of data necessarily also implies the reuse and refinement of specific machine learning models. For the private technology companies involved in the reuse of public data, such as Palantir and AWS, the value of data reuse is deeply enmeshed with the value of recalibrated and repurposed algorithms.

As entangled configurations of algorithms and data, the parameters of machine learning models are continually being rearranged and reassembled (Amoore, 2020; Seaver, 2017; Suchman, 2007). In this sense, the use of a machine learning algorithm also entails the reuse of other entities – past encounters with training datasets, the weights and parameters of a previous iteration, the back propagation of errors and the data inputs of a trial of the system (Agostinho, 2018). Consider, for example, how the neural networks for autonomous vehicles learn to respond to situations (unknown scenarios, un-encountered dangers) that were not encountered in the training data. As neural networks become increasingly adept at flexibly adapting to new situations, they are able 'to reuse the latest models and their building blocks' to address

complex problems such as speech and object recognition (Bengio et al., 2021). Bengio et al. also suggest that one of the fundamental questions in computer science is 'how could we endow neural networks with the ability to adapt quickly to new settings, by mostly reusing already known pieces of knowledge?' (2021: 63). In the practice of contemporary computer science, the idea of reuse far exceeds the repurposing of data and advances a logic of reuse that facilitates generalization to new problems based on the entangled fragments of data. In short, for machine learning, reuse is a crucial idea because it enables the generative use and application of partial knowledge (patterns, clusters, parameters) to unknown and uncertain situations.

It is worth reflecting on the political implications of reusing a machine learning model from the perspective of computer science, wherein similar computational projects can be redeployed in different domains. 'Rather than train a new model from scratch for a new application', writes Kelleher, 'we would rather repurpose models that have been trained on a similar task' so that 'it is possible to reuse the early layers of pre-trained CNNs across multiple image processing projects' (2019: 236). In this vision of the reuse of layers within a convolutional neural network, the pre-trained model learns from its previous exposures to image datasets, such that the residue of those data, images, weights and features remain lodged within the model as it is reused in a new domain. Even as the algorithm adapts to its new use, the traces of what it has learned based on its prior application determine how it behaves in the new domain. Thus, to reuse an algorithmic model in computer science is to bridge the gap between the universal (the common task or problem) and the particular (the specific instantiation of a new application). In this computational framework, however, there is also a deeply political process of constitution and implementation. The reuse of the layers of a neural network entails the creation and distribution of representations of the world in which the algorithm is deployed.

Concluding remarks

The impetus for data reuse in governments and industries relies on linear frameworks of traceability, transparency and control. However, as this article shows, machine learning generates reuse entanglements. These entanglements are enabled by platformed economies and deep learning processes, and they give rise to new taxonomies, practices and political dynamics.

Given that machine learning algorithms are designed to benefit from reuse, to leverage and learn from every fragment of data, they also actively self-generate as reusable entities. These reuse entanglements cannot be reduced to an instrumental relationship of 'master' to 'tool', but are closer to generative relations that decide who gets what, what is valuable, what is not and how an entity becomes available for use, or not. It is within these generative reuse practices, and the infrastructures and architectures that enable them, that one may identify the politics of reuse. These perspectives on the entanglement between data, algorithms and more-than-human environments highlight the need to expand analyses of the politics of data reuse beyond the bounded assessments of 'data' and 'algorithms' toward a more enmeshed framework of reuse entanglements.

Significantly, shifting the analytics from data reuse to reuse entanglements sheds light on the limitations of forms of governance that rely on transparency and traceability, as well as the seemingly benign premises of 'open science'. To shift the focus from data reuse to reuse entanglements multiplies the points of potential critique. This is meaningful, first, because it complicates the linear and discrete understanding of data reuse that underpin emerging 'data supply chain' imaginaries, allowing us instead to understand data entities as phenomena that emerge within, and are contingent upon, the techno-scientific apparatus in which they appear. Second, and relatedly, it helps us to understand how machine learning regimes do not simply reuse individual and intact previous data, but rather decompose and recompose data into non-linear and iterative forms. And third, and finally, such insights have the potential to foster a new potential for ethics and the politics of reuse and introduce broader questions about why and how requirements related to usefulness are imposed on data, and why thinking about these matters might matter beyond mere utilitarian purposes.

Acknowledgements

The authors acknowledge the Independent Research Fund Denmark grant "AI REUSE" and the ERC Advanced Grant "Algorithmic Societies: Ethical Life in the Age of Machine Learning" for supporting their research. The article is indebted to ongoing dialogues about data sets and machine learning with research communities in Denmark, the UK and abroad, including the TechSoc Cluster (Copenhagen Business School), Digital Democracies Institute (Simon Fraser University), the AI Governance and Governmentality Seminar Series (Concordia University) and the participants in the Inference Worlds panel track (EASST 2022). We particularly thank the three anonymous reviewers for their excellent reflections as well as Daniela Agostinho, Robin Steedman, Frederik Schade and Daniel Hardt for their invaluable input on earlier iterations of this article. Finally, we would like to thank the editors of *Big Data & Society* for giving this article a home.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Det Frie Forskningsråd, ERC Advanced Grant (grant numbers 9131-00115B and 883107-ALGOSOC).

ORCID iDs

Nanna Bonde Thylstrup https://orcid.org/0000-0001-6094-2970
Kristian Bondo Hansen https://orcid.org/0000-0002-9536-6050

References

Aaen J, Nielsen J and Carugati A (2021) The dark side of data ecosystems: A longitudinal study of the DAMD project. *European Journal of Information Systems*. https://doi.org/10.1080/0960085X.2021.1947753.

Agostinho D (2018) Chroma key dreams: Algorithmic visibility, fleshy images and scenes of recognition. *Philosophy of Photography* 9(2): 131–155.

Ahmed S (2019) *What's the Use?* North Carolina: Duke University Press.

Aitken R (2017) 'All data is credit data': Constituting the unbanked. *Competition & Change* 21(4): 274–300.

Alaimo C and Kallinikos J (2020) *Managing by Data: Algorithmic Categories and Organizing*. *Organization Studies*. 42(9), 1385–1407.

Alaimo C, Kallinikos J and Aaltonen A (2020) Data and value. In: *Handbook of digital innovation*. Edward Elgar Publishing, pp.162–178.

Alden W (2018) Palantir had no policy on social media data collection prior to 2015. BuzzFeed News. Available at: https://www.buzzfeednews.com/article/williamalden/palantir-had-no-policy-on-social-media-data-collection.

Amoore L (2013) *The Politics of Possibility*. North Carolina: Duke University Press.

Amoore L (2020) *Cloud Ethics*. North Carolina: Duke University Press.

Amoore L (2022) Machine learning political orders. *Review of International Studies*: 1–17.

Andrew N, Alex HH, Emily D, et al. (2020) Lines of sight. Logic Magazine. Available at: https://logicmag.io/commons/lines-of-sight/.

Barad K (2003) Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs: Journal of Women in Culture and Society* 28(3): 801–831.

Barad K (2007) *Meeting the Universe Halfway: Quantum Physics and the Entanglement of Matter and Meaning*. Durham, NC: Duke University Press.

Barnes TJ and Wilson MW (2014) Big data, social physics, and spatial analysis: The early years. *Big Data & Society* 1(1): 2053951714535365–2053951714535365.

Bengio Y, Lecun Y and Hinton G (2021) Deep learning for AI. *Communications of the ACM* 64(7): 58–65.

Birhane A, Prabhu VU and Kahembwe E (2021) Multimodal datasets: Misogyny, pornography, and malignant stereotypes. *ArXiv Preprint* ArXiv:2110.01963.

Burke A (2019) Occluded algorithms. *Big Data & Society* 6(2): 2053951719858743.

Chollet F (2018) *Deep Learning with Python*. New York: Manning.

Christen K (2018) Relationships, not records: Digital heritage and the ethics of sharing indigenous knowledge online. In: Sayers J (eds) *Routledge Companion to Media Studies and Digital Humanities*. Taylor and Francis, London: Routledge, pp.403–412.

Cifor M, Garcia P, Cowan TL, et al. (2019) Feminist Data Manifest-No. Available at: https://www.manifestno.com/.

Coldicutt R (2018) Ethics won't make software engineering better. Doteveryone. Available at: https://medium.com/doteveryone/ethics-wont-make-software-engineering-better-f3ffeca11c2c.

Custers B and Bachlechner D (2017) Advancing the EU data economy: Conditions for realizing the full potential of data reuse. *Information Polity* 22(4): 291–309.

Custers B and Uršič H (2016a) Big data and data reuse: A taxonomy of data reuse for balancing big data benefits and personal data protection. *International Data Privacy Law* 6: ipv028–ipv028.

Custers B and Uršič H (2016b) Big data and data reuse: A taxonomy of data reuse for balancing big data benefits and personal data protection. *International Data Privacy Law* 6: ipv028.

Fountaine T, McCarthy B and Saleh T (2021) Getting AI to scale. *Harvard Business Review* 99(3): 116–123.

Fourcade M and Johns F (2020) Loops, ladders and links: The recursivity of social and machine learning. *Theory and Society*, 49(5), 803–832.

Franco E (2022) The coding Prometheus is blind – socio-technological imaginaries on GitHub. *Interculture Journal: Online Zeitschrift für Interkulturelle Studien* 21(36): 49–67. https://www.interculture-journal.com/index.php/icj/article/view/442.

Goodfellow I, Bengio Y and Courville A (2016) *Deep Learning*. Cambridge, Mass: MIT Press.

Gilder L (2008) *The Age of Entanglement: When Quantum Physics Was Reborn*. New York: Random House.

Hälterlein J (2021) Epistemologies of predictive policing: Mathematical social science, social physics and machine learning. *Big Data & Society* 8(1): 20539517211003120–20539517211003120.

Hansen KB and Borch C (2022) Alternative data and sentiment analysis: Prospecting non-standard data in machine learning-driven finance. *Big Data & Society* 9(1): 1–14.

Harvey A and LaPlace J (2021) Exposing.ai. https://exposing.ai.

Hoffmann AL (2020) Terms of inclusion: Data, discourse, violence. *New Media and Society* 23(12): 3539–3556.

Hoijtink M and Planqué-van Hardeveld A (2022) Machine Learning and the Platformization of the Military: A Study of Google's Machine Learning Platform TensorFlow. *International Political Sociology* 16(2). https://doi.org/10.1093/ips/olab036.

Kelleher JD (2019) *Deep learning*. MIT press.

Keyes O and Austin J (2022) Feeling fixes: Mess and emotion in algorithmic audits. *Big Data & Society* 9(2). https://journals.sagepub.com/doi/full/10.1177/20539517221113772.

Keyes O, Nikki S and Wernimont J (2019) The government is using the most vulnerable people to test facial recognition software. *Slate*. Available at: https://slate.com/technology/2019/03/facial-recognition-nist-verification-testing-data-sets-children-immigrants-consent.html.

Leonelli S and Tempini N (2020) Data journeys in the sciences. In: *Data Journeys in the Sciences*. Available at: https://doi.org/10.1007/978-3-030-37177-7.

Loukissas YA (2019) All Data Are Local. In: *All Data Are Local*. Available at: https://doi.org/10.7551/mitpress/11543.001.0001.

Madianou M (2020) Covid19 A second-order disaster? Digital technologies during the COVID-19 pandemic. *Social Media + Society* 6(3): 2056305120948168.

Martin A, Sharma G, Peter de Souza S, et al. (2022) Digitisation and sovereignty in humanitarian space: Technologies, territories and tensions. *Geopolitics*. https://doi.org/10.1080/14650045.2022.2047468.

Mayer-Schönberger V and Cukier K (2013) *Big Data: A Revolution that will Transform how We Live, Work, and Think*. London, UK: John Murray.

McQuillan D (2018) Data science as machinic neoplatonism. *Philos. Technol* 31: 253–272.

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. *Big Data and Society* 3(2): 1–21.

Mittelstadt BD and Floridi L (2016) *The Ethics of Biomedical Big Data (Vol. 29)*. Berlin, Germany: Springer.

Mulvin D (2021) *Proxies: The Cultural Work of Standing In*. Cambridge, Mass: MIT Press.

Narayan D (2022) Platform capitalism and cloud infrastructure: Theorizing a hyper-scalable computing regime. *Environment and Planning A: Economy and Space* 54(5): 911–929.

Pasquetto IV, Randles BM and Borgman CL (2017) On the reuse of scientific data. *Data Science Journal* 16; 8: 1–9, https://doi.org/10.5334/dsj- 2017-008.

Peng K (2020) Facial recognition datasets are being widely used despite being taken down due to ethical concerns. Here's how. Freedom to Tinker. Availabe at: https://freedom-to-tinker.com/2020/10/21/facial-recognition-datasets-are-being-widely-used-despite-being-taken-down-due-to-ethical-concerns-heres-how/ (accessed on July 15 2022).

Plantin JC, Lagoze C and Edwards PN (2018) Re-integrating scholarly infrastructure: The ambiguous role of data sharing platforms. *Big Data and Society* 5(1). https://doi.org/10.1177/2053951718756683.

Radin J (2017) 'Digital Natives': How Medical and Indigenous Histories Matter for Big Data. *Osiris* 32(1): 43–64.

Raji ID, Gebru T, Mitchell M, et al. (2020) Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp.145–151).

Ribes D, Hoffman AS, Slota SC, et al. (2019) The logic of domains. *Social Studies of Science* 49(3): 281–309.

Sands EG (2018) How to build great data products. *Harvard Business Review*.

Schneier B (2016) Data is a toxic asset. *CNN Blog*, Available at: https://edition.cnn.com/2016/03/01/opinions/data-is-a-toxic-asset-opinionschneier/index.html.

Schwarzkopf S (2020) Sacred excess: Organizational ignorance in an age of toxic data. *Organization Studies* 41(2): 197–217.

Seaver N (2017) Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society* 4(2): 2053951717738104.

Slota SC, Hoffman AS, Ribes D, et al. (2020) Prospecting (in) the data sciences. *Big Data and Society* 7(1). https://doi.org/10.1177/2053951720906849.

Stevens N and Keyes O (2021) Seeing infrastructure: Race, facial recognition and the politics of data. *Cultural Studies*: 1–21.

Suchman L (2007) *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge, UK: Cambridge University Press.

Sutherland T. (2019). The Carceral Archive: Documentary Records, Narrative Construction, and Predictive Risk Assessment. *Journal of Cultural Analytics.* https://doi.org/10.22148/16.039.

Sutherland T (2021) Remains. In: Thylstrup NB, Agostinho D, Ring A, D'Ignazio C and Veel K (eds) *Uncertain Archives: Critical Keywords for Big Data*. Cambridge, Mass: MIT Press, pp.433–442.

Thatcher J, O'Sullivan D and Mahmoudi D (2016) Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space* 34(6): 990–1006.

Tenopir C, Dalton ED, Allard S, et al. (2015) Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PloS one* 10(8). https://doi.org/10.1371/journal.pone.0134826.

Thylstrup NB (2019) Data out of place: Toxic traces and the politics of recycling. *Big Data and Society* 6(2). https://doi.org/10.1177/2053951719875479.

Thylstrup NB (2022) The ethics and politics of datasets: Deleting traces and encountering remains. *Media, Culture & Society* 44(4), 655–671. https://doi.org/10.1177/01634437211060226.

Thylstrup NB, Agostinho D, D'Ignazio C, et al. (2021) *Uncertain archives : Critical keywords for big data*. Cambridge, Mass: MIT Press.

United States (2021) *Interim review of the national security commission on artificial intelligence effort and recommendations : Hearing before the subcommittee on intelligence and emerging threats and capabilities of the committee on armed services house of representatives one hundred sixteenth congress second session hearing held : september 17* 2020. U.S. Government Publishing Office. Availabe at: https://purl.fdlp.gov/GPO/gpo159185 (accessed on November 23 2022).

Van De Sandt S, Dallmeier-Tiessen S, Lavasa A, et al. (2019) The definition of reuse. *Data Science Journal* 18(1): 1–19.

Van Noorden R (2020) The ethical questions that haunt facial-recognition research. *Nature : International Weekly Journal of Science* 354–358. https://doi.org/10.1038/d41586-020-03187-3.

Verhulst SG (2020) Unlock the Hidden Value of Your Data. *Harvard Business Review*.

Wilkinson MD, Dumontier M, Aalbersberg Ij J, et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3(1): 1–9.

Winkler CE and Berenbon RF (2021) Validation of a survey for measuring scientists' attitudes toward data reuse. *Journal of the Association for Information Science and Technology* 72(4): 449–453.

Zimmerman AS (2008) New knowledge from old data: The role of standards in the sharing and reuse of ecological data. *Science, Technology, & Human Values* 33(5): 631–652.

Zook M, Barocas S, Boyd D, et al. (2017) Ten simple rules for responsible big data research. *PLoS Computational Biology* 13(3): e1005399.

Zuboff S (2019) *The age of surveillance capitalism*. New York: Public Affairs.