Original research article

BIG DATA & SOCIETY

Big Data & Society
October–December: 1–15

DOI: 10.1177/20539517251396092
journals.sagepub.com/home/bds
Sage

Feeding the machine: Practitioner experiences of efforts to overcome AI's data dilemma

Jo Bates[1], Monika Fratczak[2], Helen Kennedy[2], Itzelle Medina Perea[1] and Erinma Ochu[3]

Abstract

This paper examines the human implications of AI's 'data dilemma' in three different and contrasting sectors: pharmaceuticals, higher education, and the arts. The 'data dilemma' refers to the challenge of obtaining sufficient and suitable data to effectively train AI algorithms. The research, conducted in the UK, involved interviews, focus groups, and observations with 65 practitioners employed across these three sectors. The findings reveal that addressing the data dilemma often involves practitioners being pressured to generate data for AI, either passively in the context of data extractivism or actively by engaging in new forms of data production. We explore how this pressure to 'feed the machine' manifests differently in each sector, and how modes of resistance to these emergent data practices vary across sectors. We observe that the push to resolve the data dilemma is fundamentally driven by capitalist and technological solutionist values; values that often conflict with those of practitioners who are expected to adapt their practices in the service of AI-driven capitalism. We conclude with a call for exploring different approaches to AI development that align with alternative value systems.

Keywords

data inputs, AI, machine learning, data dilemma, data work, practitioners

Introduction: AI's data dilemma

As AI permeates our lives, the quality of data available to train models becomes increasingly crucial. However, in the current era of data abundance, a paradox emerges: while we have access to more data than ever before, much of it is of poor quality or inappropriate for training models, leading to AI models that are biased, inaccurate, and ultimately flawed – a phenomenon termed AI's 'data dilemma' by the UK's Turing Institute (2024). Research that critically examines how practitioners experience the integration of predictive and generative AI algorithms in actual contexts of practice has focused on themes such as: negotiations between different groups of practitioners (e.g., Passi and Sengers, 2020; Stalph, 2020); how particular jobs are evolving with the adoption of algorithmic technologies (e.g., Chan et al., 2022; Dencik and Stevens, 2021); and, the impact of automating professional work (e.g., Turja et al., 2022; López Jiménez and Ouariachi, 2021). Little attention has been paid to the emerging pressures felt by skilled practitioners around increasing the amount of quality data that can be fed to hungry AI algorithms. Yet, this was a key theme in our empirical research examining the beliefs, values, and emotions of practitioners about the integration of different types of AI into work practices in the pharmaceutical, higher education and arts sectors.

While research in critical data studies (CDS) has identified a deepening 'desire for numbers' (Kennedy, 2016) across different sectors, the focus of that research has been on understanding the epistemic desire for actionable

[1]School of Information, Journalism and Communication, University of Sheffield, UK
[2]School of Sociological Studies, Politics and International Relations, University of Sheffield, Sheffield, UK
[3]Faculty of Arts, Creative Industries, and Education, University of West England, Bristol, UK

Corresponding author:
Jo Bates, The Wave, 2 Whitham Road, Sheffield S10 2AH, UK.
Email: Jo.bates@sheffield.ac.uk

quantitative *outputs* that result from some form of analytics, whether descriptive or predictive. This same orientation is also present in, for example, Science and Technology Studies research on cultures of prediction, where attention has been on how numerical predictions contribute to the stabilisation and legitimisation of knowledge claims (Fine 2007; Heymann et al., 2017). Similar to studies on practitioners' experiences of AI integration, this research on desire for predictive outputs does not address the growing need for large amounts of quality data with which to feed the systems that produce these numbers, among other outputs. While it is recognised that datafication in general brings new forms of 'data work' for practitioners (Pine et al., 2022; Jarke and Buchner, 2024), this research hasn't addressed such data work in the context of AI's data dilemma. This matters because a focus on data *inputs* brings our attention to how practitioners are experiencing, and in some cases resisting, growing pressures to address AI's data dilemma and ultimately meet the needs of contemporary data capitalism. The focus on data inputs is therefore an important consideration for understanding the impact of, and engagement with, AI adoption in the workforce.

In this paper, we therefore address the 'cultures of data practice' that shape how practitioners experience this pressure to 'feed the machine' – a term used by Muldoon et al. (2024) in their research on the precarious global workforce enabling AI systems, but which we use more broadly to consider the labour of skilled practitioners that is required to sustain both AI machines and the capitalist system within which they are embedded. We draw on empirical research conducted in the UK, which consists of interviews, focus groups and observations with 65 practitioners in the pharmaceutical industry, Higher Education and arts practice. Our analysis of this qualitative data leads us to question the assumption that the machines can be put to work on an abundance of available data as sometimes imagined in hyped accounts of AI's transformative potential. We argue that, in some contexts of practice, efforts to integrate AI algorithms into workflows leads to increased pressures on practitioners to keep feeding the machine, whether actively through engaging in new forms of data production practices or passively as tech companies engage in the extractivist practice of harvesting practitioners' online works. Comparison across the three contexts that we examined also leads us to argue that pressure to engage in feeding the machine appears to be stronger the more embedded a sector is within capitalist values, i.e., profit motive, self-interest, efficiency and competition, while resistant cultures are more observable the further practitioners are situated from capitalist value systems.

This paper begins by discussing related research on the topic of data production from CDS and related fields, before discussing our methodological approach. We then present three narratives from each of our cases which identify pressure points around 'feeding the machine' in that context of practice. Finally, we discuss our findings and how they contribute to literature on practitioners' experiences of predictive AI integration into their workflows.

The production of data inputs

While the datafication processes in contexts such as surveillance and social media are well documented and critiqued (e.g., Benjamin, 2019; van Dijk, 2013), less attention has been paid to data inputs as machine learning and other AI techniques have been 'generalised' (Mackenzie, 2015) into sectors that are not as data rich. Yet, as Thylstrup (2022) argues in her call for more 'critical dataset studies', the 'global push for machine cultures has given rise to an increasing demand for data' (p. 655). This push has led to more questions being asked about the datasets that are available to fuel AI advances (e.g., McKendrick, 2024). For example, a recent Turing Institute (UK) seminar brought together speakers from across UK government bodies and the British Computer Society to discuss data quality challenges in the context of AI adoption in each of their organisational contexts (Turing Institute, 2024). However, there is limited exploration of what this 'data dilemma' (Turing Institute, 2024) means for practitioners on the ground in different contexts of practice. In particular, little attention has been paid to the expectation and pressure on practitioners to enable the feeding of AI machines with more and better data.

Exceptions in the CDS and STS literatures that begin to address data input challenges for practitioners can be found in the context of health data research. Hoeyer (2019) introduces the concept of 'promissory data' to explain how burdensome data collection initiatives aimed at personalised medicine work to responsibilise patients and professionals to engage in data work geared towards 'a promise for an unspecified future' (p. 1). Medina Perea (2021) observes the related pressure that health researchers who have absorbed the 'big promises of big data' place on themselves to improve the quality and quantity of patient health data feeding into university research labs. Further, Choroszewicz (2022) describes the emotional labour of 'care' in data practices that enable health data re-use. The theme of care in the context of datafication is also picked up by Jarke and Büchner (2024) who address the ways that professionals in other public service contexts are often drawn into new 'data care arrangements…in addition to and alongside their existing tasks'. This research serves to draw attention to the ways that practitioners in public service contexts are increasingly drawn into new forms of data work, in some cases of their own volition and motivated by a commitment to improving data-driven practices in their working context. However, in these examples, the pressures described, particularly in relation to AI integration in specific contexts, receive little attention, thus our understanding of them is limited in scope and depth.

While there has been limited research on the pressure to feed algorithmic systems with more and better data, critical research on data more generally informs our understanding of the cultures that are shaping data inputs. For example, CDS scholars have addressed the epistemic issues in the datasets that feed algorithmic cultures. It is widely recognised that 'raw data is an oxymoron' (Bowker, 2008; Gitelman, 2013), and that data – and thus datasets – are created and shaped by people embedded in complex socio-cultural settings that influence this production process (e.g., Kitchin, 2014; Bates et al., 2016). We therefore know that beliefs and values, as well as material contexts, play a significant role in the production and use of data as inputs. We also know that these data practices have an affective component (Kennedy and Hill, 2018). For example, research has identified how the production of self-tracking data can be associated with feelings of anxiety or frustration (e.g., Lupton, 2019; Pantzar and Ruckenstein, 2014). Little research, however, considers how these beliefs, values and, importantly, emotions, as well as the material contexts these develop and evolve within, work together to constitute what we call 'cultures of data practice', and which we use to guide our own empirical focus as expanded in the following section.

In some cases, these cultures of data practice result in biases in datasets. Of most interest in the CDS literature have been the systematic socio-cultural biases that are often baked into the datasets used to train algorithmic systems resulting in discriminatory outputs (e.g., Barocas and Selbst, 2016; Benjamin, 2019; Noble, 2018). Sometimes these biases are the result of 'gaps' in pre-existing datasets that result from dominant belief and value-systems that are implicit in the way that practitioners produce data inputs and resultantly shape the constitution of datasets. For example, facial recognition systems that were trained on primarily lighter-skinned faces (Buolamwini and Gebru, 2018). The issue of gaps in existing datasets has been of particular interest in the data activism literature. For example, Gabrys' (2016) 'Just Good Enough' data is a classic example of a study examining how citizens, informed by their beliefs and values, worked to fill a gap in the official data about their local environment by producing their own air quality data to feed into policy-making processes. Other examples include the work of Currie et al. (2019), who examine the work of 'missing data' projects run by activists in Kosovo and the USA, as well as other activist groups such as Data for Black Lives and the Interactive Map of Femicides in Mexico. These contributions all demonstrate the ways in which alternative data production practices have been used to counter the dominant values and beliefs of racialised capitalism that have shaped datasets in ways that reproduce and reinforce environmental harm, structural racism, ableism and misogyny. However, within the CDS and related research there has been less focus on other, less obviously politically charged, domains, and less attention on how such values and beliefs influence how practitioners engage with the growing data input requirements of AI technologies.

Other research has examined exploitative practices in the production of data underlying algorithmic systems. For example, we see growing attention paid to labour exploitation in the production of annotations for datasets that are used in the training of machine learning and similar systems. Researchers have examined the crowdwork infrastructures that produce these datasets, uncovering the exploitative labour conditions of the workers involved (e.g., Gray and Suri, 2019; Graham et al. 2017). This research overlaps with that on content moderation practices, which observes the emotional and psychological impacts on workers who create data labels for graphic and hateful social media content (e.g., Roberts, 2019). This body of research raises questions about the hidden systems of oppression that underlie the seemingly polished exterior of algorithmic systems. However, this existing literature tends to focus on precarious labour and marginalised populations, rather than the professional labour of, for example, scientists, educators and artists that is leveraged in the production and management of data to feed AI machines.

In this paper, we bring together these observations from the existing research. We also recognise that the demands, practices and concerns identified in existing research are shaped by cultures of data practice, constituted of values, beliefs and emotions that develop and evolve within material contexts. Importantly, we argue that we should understand these cultural dynamics holistically rather than separately. We also note that there is currently little critical exploration of data input practices in the context of professional labour and in less politically charged settings. There is therefore little understanding of how professional cultures of data practice are experiencing these growing demands for data in sectors that are experiencing efforts to integrate AI based techniques, such as predictive machine learning, into working practices. Understanding these issues is important because they raise questions about the potential human implications of AI in everyday practice, and raise further challenges for decision makers about what it means to adopt AI responsibly.

Understanding AI integration in three contexts of practice: Research methods

We collected the empirical data that we draw on as part of the Patterns in Practice (AHRC) project, through which we aimed to understand how practitioners' beliefs, values and emotions interact to shape how they engage with different and contrasting types of applied machine learning and related forms of AI. By practitioners, we mean people who work in professions that involve a high level of training

and skill. We were interested in engaging with practitioners who do the computational and technical work behind machine learning systems and their implementation, as well as those whose profession means they engage with the outputs of such systems in their work and those that manage this integration of AI into workflows.

Our project focused on practitioners in three distinct and contrasting sectors: pharmaceuticals, higher education and the arts. The specific AI applications we focused on were machine learning for drug discovery, learning analytics in universities, and artists' use of machine learning and related techniques in their practice. We selected these three sectors as examples of scientific applications involving no data about people, social applications involving data about people, and creative applications that may or may not involve data about people. We also selected them due to their contrasting business models, with pharmaceuticals being a highly competitive and profitable private sector, higher education a regulated and marketised public service, and the arts a largely independent freelance based sector. Through our selection of cases with contrasting material conditions shaping AI adoption and practitioner experiences, we aimed to capture a nuanced understanding of what was happening across different parts of the economy, as well as identify general themes that cut across these very different contexts of practice. The theme of this paper – feeding the machine – is one such cross cutting theme.

We recruited participants differently depending on the sector and how it was organised. For the pharmaceuticals case we worked with one multinational pharmaceutical company where we recruited people across three projects working towards new forms of machine learning integration in the small molecules part of the industry (Interview codes Px). For this case we selected a part of the sector that has a long history of working with predictive algorithms for drug discovery and drew on professional networks in the informatics space to gain access. For the higher education case we examined efforts to integrate predictive learning analytics into university processes with Jisc (a non-profit organisation in the UK that supports IT and digital provision in further and higher education; Interview codes EJx), two English universities with different experiences of learning analytics adoption (Interview codes EU1.x and EU2.x), and also practitioners in the wider Jisc learning analytics network (Interview codes EJNx). For the arts case, reflecting the much more independent nature of the sector, we recruited practitioners who were experimenting with different types of AI techniques through desk research and snowball sampling (Interview codes Ax). All the computational techniques and technologies practitioners were reflecting on (i.e., machine learning, predictive learning analytics) can be defined as forms of narrow AI, that is AI that is programmed to perform a specific task such as predicting student achievement or generating content.

Across the three cases we conducted a mixture of interviews, focus groups and observations with 65 UK-based practitioners. This involved 18 interviews and 2 focus groups in the pharma case, 33 interviews and 4 focus groups in the Higher Education case and 14 interviews and 3 focus groups in the arts case. Focus groups each had between 3 and 6 participants who had already been interviewed, and topics focused on emerging themes coming out of interview analysis. We also observed staff and network meetings in the pharma case (two staff meetings) and education case (two staff meetings; one network meeting), as well as a number of arts events (one festival; one exhibition; three panels). We used these observations to sensitise ourselves to practitioners' contexts of practice, rather than for formal analysis. Our data collection started in Summer 2022 and finished in Summer 2023. We initially analysed data from interview and focus group transcripts using thematic analysis (Braun and Clarke, 2006) to draw out the key topics that emerged when practitioners talked about their beliefs, values and emotions in relation to AI integration into their workflow. Transcripts were inductively coded in Nvivo by two team members, prior to extraction of the codebooks for each case for use in full team workshops aimed at theme generation. In these workshops we identified both sector specific and cross cutting themes. Cross cutting themes offering original insights included: 'Feeding the machine' (the focus of this paper), 'Surprise', 'Tactics of resistance' and 'Human-machine collaboration', each of which is developed in a separate paper. Once themes had been identified, we re-analysed transcripts to check they were coherent themes and undertake a close critical reading in relation to each identified theme. In this paper, we present findings from one of these themes – 'feeding the machine', which emerged clearly in the data collected across all three cases as participants reflected on their experiences and struggles with AI integration. Our close critical reading around this theme involved exploring stories about 'pressure points' that arose in our data around the theme of 'feeding the machine'. This included analysing the context that was leading to this pressure to form and how differently situated practitioners were experiencing it. Given our focus was on the beliefs, values and emotions of practitioners, our focus is on this, rather than a detailed exposition of data production and use. Ethical approval for the study was gained from University of Sheffield.

Findings: Feeding the machine

The challenge of 'feeding the machine' was a key theme in each of our cases, yet it played out in different ways as a result of the different forms of AI being adopted and how practitioners in the sector were situated in relation to the forces of capitalism and/or the neoliberal regulatory state. In this section we present our detailed findings on the feeding the machine theme from each of the three cases. We

begin with the pharmaceuticals case, before moving on to higher education and finally the arts. In each section, we narrate the story of the 'data pressure point' identified in that case, beginning with an explanation of the context that has led to pressure to form from the perspective of practitioners, before going to explore their perception and feelings about solutions that have been adopted to address the data dilemma in that context. A comparative table 1 is provided at the end of the findings section.

The pharma industry's 'bad data' problem

Adoption of predictive models in pharma

The use of predictive models has a long history in the pharmaceutical industry, with waves of interest stretching back a number of decades. However, while in previous decades computational work was relatively 'peripheral', today it is seen as 'central to the drug discovery process' (P11) and advances in data-hungry techniques such as deep learning have begun to disrupt the industry. In the part of the sector we examined, the data used to train models is chemical 'fingerprint' data and data resulting from experiments in wet labs, rather than any personal data. Predictive techniques are used on such data to better understand e.g., how compounds are likely to interact and what might be toxic for humans, and the results inform which ideas should be taken forward in costly experiments. Practitioners that we spoke to across different roles, while sceptical of the current hype around AI for drug discovery, believed there was significant potential in applying predictive technologies in the longer term.

> We recognise that the number of possible molecules you can make is so large that no human can ever make sense of it, or think of all of it, but potentially a computer can. (P07)

A key driver for adopting predictive technologies is the low success rates experienced within the drug discovery pipeline, and the financial cost of this challenge. As one scientific manager we spoke to explained:

> We have awful attrition. So, if you put 100 drugs – by the time you get to test it in a human… you're only going to end up with five out the other end… you're really paying for the ninety projects that failed rather than the ones that succeed. (P11)

It can also take 'twenty, twenty-five years' for some drugs to get to market, by which time 'the patent has already expired' (P10). Participants working in the pharma industry perceived a growing drive from leaders to adopt predictive models that have the potential to 'speed up the entire process' (P15) and enable experiments to 'fail[…] less often' (P11). This was based on a belief that doing so would enable the firm to remain competitive in a fast moving and pressurised sector, subject to the logic of acceleration (i.e., 'the setting-in-motion of the material, the social and the cultural world at an ever increasing speed') that Rosa et al. (2017: 58) argue is a 'shared essence of all capitalisms'.

> Everybody's objective is to speed up that process… there's always some debate about the balance between speed and quality…that goal is shared by everyone. (P02)

The values underlying this shared desire to accelerate the process varied between the profit driven values of the company as a whole and the more altruistic values of many practitioners whose primary driver was related to improving people's health.

> Knowing that some of the projects that you work on might lead to a drug that might save somebody's life or make it better really motivates me…it's nice to know that something that you really enjoy doing can have such a positive impact. (P06)

However, while the values driving practice sometimes differed, they all fit with the ultimate objective to speed up the process, and no participants voiced concern about any tension between the somewhat contrasting values of altruism and profit underlying this objective and driving the resultant interest in AI adoption.

The data pressure point

Practitioners believed that there have already been some significant, and for some surprising, AI advances in the field, for example DeepMind's groundbreaking AlphaFold which predicts protein structures (P01, P02, P03, P07, P12). However, many noted that underlying this success was many years of data curation activity:

> AlphaFold depends on a lot of scientific data, and it was because the scientific community curated that data for the last twenty, thirty years, that AlphaFold is successful now… we should not forget what is the underlying structure. (P16)

As people are beginning to observe in other industries and the public sector (McKendrick, 2024; Turing Institute, 2024), such a rich source of data is not readily available for many drug discovery problems:

> For us the limitation isn't the machine learning or the resources…the limitation is having the data to work with to do that. (P10)

While companies such as the one we worked with have significantly more data than other stakeholders in the pharmaceutical sector such as university labs and AI startups (P01), the volume of data was perceived to be small when compared to other industries such as social media or finance:

Table 1. Comparative analysis of how the AI data dilemma is playing out in each AI use case.

Sector Application domain	Pharma Scientific	Higher Education Social	Arts Creative
Use case	Small molecule cheminformatics aimed at informing wet-lab experimental work in drug discovery e.g. predicting toxicity and interactions of novel compounds.	Learning Analytics systems that e.g. predict students' likelihood of success based on engagement	Use of machine learning and related method to develop artworks involving e.g. image, sound etc, including through the use of GenAI
Key business models relevant to the case (participants in **bold**)	**Multi-national public company with shareholders in a highly competitive sector** Technology firms such as DeepMind entering the sector with innovations such as AlphaFold	**Government regulated and highly marketized English universities**, with growing regulatory pressure from Office for Students for data-driven accountability **Non-profit organisation Jisc which supports universities with IT and digital provision – supplier of learning analytics product** Private companies in the EdTech sector – suppliers of Learning Analytics products to universities	**Independent freelance artistic practitioners** Big Tech firms such as Open AI developing GenAI products such as DALL.E
Data type	Chemical and experimental data – no data about people. Data is a mix of that widely available in the industry and data generated from confidential experiments conducted within the firm	Data about students' activities e.g. on VLE and registration systems, and notes from staff-student interactions	Data harvested from online sources e.g. online images, sounds, and text uploaded to the internet by individual users and organisations (Big Tech) Data created by artists themselves for artistic purposes (arts practitioners)
History	Long history of use of predictive algorithms in cheminformatics with more recent work exploring new, advanced models e.g. deep learning	Student engagement data relatively basic and ad-hoc until recent years e.g. was kept in departmental or individual spreadsheets if collected. Adoption of centrally managed learning analytics tools is fairly recent, over the last decade.	Long history of artists experimenting with AI and machine learning, dating back to 1960s. Text to image generators have evolved over the last decade.
Key problem being addressed by AI adoption from perspective of key decision makers	Attrition – most ideas for drugs fail before they get to market which is expensive. Firms need to predict better which compounds have the most likelihood of success, so they can invest there.	Student retention rates need to be monitored (and in some institutions increased) to meet the regulatory demands of the Office for Students.	No specific problem being solved by GenAI – possibly a solution in search of a problem. N/A with regard to arts practice, not problem-solution oriented.
The data dilemma	Lack of the right type of data to train AI models. Existing datasets are unbalanced with not enough data on bad molecules from failed experiments as this was previously disposed of.	Accuracy and completeness of data held by institutions about students is not good enough to train predictive models	GenAI firms need significant and growing amounts of data to train models
Practices to address data dilemma	Data production – designed a project to run experiments that would fail to generate more data on bad molecules.	Data production - time-consuming task of recording data about any interactions with students	Data harvesting - GenAI firms harvest e.g. images from the internet, with no consideration of intellectual property or labour rights

(continued)

Table 1. Continued.

Sector Application domain	Pharma Scientific	Higher Education Social	Arts Creative
	Resulted in significant loss of motivation and frustration.	adding to teachers' stress and workload	Artists indirectly address this data dilemma by offering critique of Big Tech's extractive solution to its data dilemma and frequently using open source tools and their own curated datasets in their own practice.
Values	Efficiency/speed to deliver profit and compete (firm as a whole) Efficiency/speed to deliver altruistic ends (some individual practitioners) Reasonable autonomy and creativity in work	Ethical values reflected in concerns about: • reliability of predictions at individual student level – Jisc • universities capacity to respond adequately to students flagged as a concern – Jisc • using demographic data in predictions – Jisc and universities Careful - slow and steady approach	Arts practitioners: • Accountability, fairness, responsibility and transparency • Freedom to critique and experiment • Small, careful, ethical practices, often their own data, rather than relying on big datasets and large-scale appropriation of others' work • Environmental sustainability, particularly in the global south.
Beliefs	AI is being hyped in the sector Drug discovery needs to speed up Sector leaders are increasing pressure to adopt AI to deliver efficiencies Sustained data curation is essential to deliver significant AI outcomes e.g. in case of AlphaFold	Changes in the regulatory environment (Office for Students), with an emphasis on retention, are a key driver for institutional adoption of LA in some universities. Universities also need to retain students due to financial issues as students bring money. EdTech firms have a different value system to above, and are using demographic data to inform predictions about students' success which poses ethical concerns Models don't capture and integrate interventions based on predictions – undermines models and a possible future data dilemma.	AI is being hyped in the sector Data extractivism for GenAI is a serious problem for the arts sector Arts practitioners' values are always ultimately entangled within the capitalist system and extractivist industry There's no straightforward solution to GenAI data extractivism Small data can be understood as a form of resistance (although critiqued by one person) Adhering to personal values can help to navigate challenges
Emotions	Rare surprises e.g. success of AlphaFold Frustration at the lack of appropriate data for training models Worry about grand promises of AI made by some in sector Boredom, soul destroyed, unmotivated of practitioners working on data generation project Mixed feelings - excited but sceptical	Frustrations about extra data production tasks Stressful working environment Worry that educators start to service the machine rather than support students	Enthusiasm about applying AI tools in their practice Scepticism about AI hype Anger about data extraction Pressure to keep up with technological advances, and frustration about this at times Mixed feelings – Uncertainty, some optimism, but more pessimism

> I could probably put all of [our company's] screening data onto a standard USB stick…an eight gigabytes, sixteen gigabyte disc, you know? It's not big, big data as we understand it… we often only have a hundred datapoints. You're not going to put that into a deep learning model. (P03)

A large proportion of practitioners that we spoke to across different roles recognised this lack of the right type and amount of data as a challenge for the sector (e.g., P01; P02; P05; P10; P11; P14; P15). This could be a 'source of frustration' (P01), and caused some to worry if they saw colleagues and other companies in the sector making promises to deliver AI driven results when they believed the projects would 'be data limited in a lot of cases' (P14).

> these deep neural networks…those methods work really well when you've got huge amounts of data. We don't always have huge amounts of data in our industry, so we have to be very careful that people just don't say, right, you need to apply this when it's not appropriate for what we're doing. (P02)

Tackling the 'bad data' problem

This challenge for the sector has led to a number of efforts to generate and better manage data to 'feed[…] the machine learning' (P11). One participant referred to this type of work as the 'plumbing' work in the AI race (P12), in that the labour involved is essential, but often undesirable and undervalued. Here, we consider two such efforts: data auditing and data production.

A key challenge has been to understand what historical experimental data the firm already has, and what are the gaps and problems with it if it is to be used for predictive modelling, something that was not necessarily anticipated when it was generated. One significant gap perceived in the existing data is data about 'bad molecules':

> Often, we end up with lots of [data about] very good molecules because we have a process that's been honed for years, and actually for the algorithms we need good molecules and bad molecules. We need to have a balanced training set. (P01)

A project was set up internally in the company to try to address this perceived 'bias […] towards good data' (P09) in the existing historical data. The aim of this project was to run experiments as instructed by an algorithm in order to increase the variety of data – including negative results – in the database.

While the task of auditing the data and identifying issues and gaps had been experienced as 'quite enjoy[able]… interesting' (P10), that was not the case for those who were put to work in the lab to generate 'bad data' to fill the gaps. The medicinal chemists in such roles expressed concerns that their labour and insights were devalued in the approach taken by the project.

> [I]t was basically ninety-nine percent of the time in the lab… absolutely soul destroying. It was boring as hell…I think it just kind of crushed any sort of creativity that you can have in your job. (P05)

> So, it was quite a rigid process …I gave up questioning things – actually, that's a good point. I just gave up because it wasn't worth the fight in that process. (P04)

More so than other data 'plumbing' work, the project aimed at producing experimental data about 'bad molecules' was experienced negatively by those having to follow the algorithms to feed the machine with better data. Their description of their experience is suggestive of feeling a total loss of agency to shape the process or make decisions, and in the corporate context they worked in little scope for resistance other than sharing how they felt with colleagues:

> I've spoken to multiple people who you will speak to over the next couple of days, and I've fed this back to them, it's soul destroying. (P05).

Their experience on this project contrasted with the values they held in relation to scientific work. Their reflections indicated that it was important for them to feel that their work was enjoyable, to some extent autonomous, and enabled them to use their creativity. Perhaps if the project had been a significant success in the push to train predictive models with better data our participants would have felt differently in hindsight, but there was no sense of belief that this had happened: 'And I wouldn't say that [it] was a huge success' (P05). This lack of resolution for the 'bad data' problem contributes to the mixed emotions some have about AI adoption in the sector:

> Though I'm excited by it, I'm still sceptical because I'm still worried about…the lack of bad data, whether we can get enough bad data. (P09)

Despite claiming to have spoken to many people about their negative experience on the project, only one other participant who was not put on the project, a computational chemist, reflected on the significant challenge that was faced by the sector in being able to motivate people to engage in this type of work:

> Plumbing isn't sexy…so therefore this is work nobody wants to do. Actually, this is a big problem…That is a challenge because everybody wants to be motivated. If you don't find a way of motivating the people who create the really high-quality input for the models then we won't succeed. (P12)

In summary, the practitioners we engaged with in the pharmaceutical sector clearly experienced the pressure to accelerate processes in the context of AI-driven capitalism, and

this drove them to find solutions for their data dilemma with the hope of remaining competitive within a fast-evolving industry. In some cases, this meant skilled practitioners being drawn into 'soul destroying' work to produce new 'bad' data in an effort to rebalance datasets and to resolve the dilemma, highlighting the potential human costs of feeding the machine.

Sorting out the data in higher education

Adoption of learning analytics in English higher education

While data about student engagement has been used in some Higher Education (HE) institutions for many years, until relatively recently such data tended to be held locally in departmental and personal spreadsheets (EU1.03). Many of our participants from the HE sector believed this had changed in recent years, with the increasing adoption of centrally managed learning analytics tools. While many of our participants saw some value in these tools for supporting students, a key driver for the widespread adoption of learning analytics in the English HE sector was believed to be the changing regulatory and funding environment. Specifically, the Office for Students (OFS) which was established in 2017, and which is the new quality assessment authority for English Higher Education. Similar to arguments put forward by others (e.g., Williamson, 2019), many believed that this transition had placed new pressures on HE institutions, as the OFS has driven an agenda of increasing data-driven accountability:

> It's all, prove to us that your students are attending, prove to us that your students are continuing, prove to us that your students are achieving. (EU1.08)

Universities are now required by the OFS to use data to demonstrate academic achievement, student retention and graduation rates, and post-graduate employability (Office for Students, 2024a). These requirements are in addition to existing data requirements on institutions as part of the annual Higher Education Statistics Agency (HESA) submission, and growing demands on institutions to monitor and evidence engagement of international students by the Home Office's UK Visa and Immigration (UKVI) division (Gov.uk, 2024). Failure to comply with these requirements can result in a HE institution being investigated by OFS, as well as possible fines, funding and licensing implications.

> The Office for Students are now requiring universities to prove that they're actively managing their continuation, completion and progression…if they're not up to a standard or to a benchmark then there may be, you know, a knock on the door, measures that have to be taken to improve that institution. (EJ05)

Within a wider context of financial challenges for English higher education resulting from the tuition fee funding model and home students' fees freeze (OFS, 2024b), participants believed that learning analytics solutions were often viewed by institutions as a way of monitoring and intervening in students' engagement:

> So that perhaps is a rather non-academic, unfashionable answer to do with money, and universities needing to make sure they retain students in order to meet their spending needs and requirements. (EU1.06)

It is within this wider context that non-profit Jisc introduced a HE learning analytics solution (also called Jisc by interviewees), which included an optional predictive component, in 2017. The development of this predictive tool responded to a perceived demand from the sector emerging from a consultation (EJ06).

> When we started talking to institutions, they were talking very much about wanting… AI and predictive models for learning analytics… being able to use AI to identify students who were at risk of not achieving. And also, they were interested in models to help students to understand how they could achieve better. (EJ04)

The 'data quality' pressure point

However, after a number of years the development of the predictive component was abandoned by Jisc. There were various values-driven reasons behind this decision including ethical concerns about the reliability of predictions at the individual student level (EJ04) and institutions' capacity and legal liability to respond adequately to students flagged as a concern (EJ02). However, a more significant issue was believed to be the quality of the data underlying the predictions; a challenge which has resulted in efforts to improve the data that is fed into learning analytics systems. As one Jisc practitioner explained:

> the accuracy of the data and the completeness of the data held in the institutions was often not good enough. (EJ04)

Participants reported that it was not uncommon for universities that explored the possibility of adopting learning analytics to find their beliefs about their data quality challenged:

> We thought our data was quite good… but when we started to look at it, we realised how unclean our data was, even from a HESA return perspective. (EJ02, formerly UKHE institution)

As a result, some participants believed English universities were not yet ready to adopt reliable predictive analytics.

> we have to have a realistic thought about the journeys that people need to go through in terms of data quality at all of the institutions. (EJ03)

For them, it was more aligned with their values to address the issues with the datasets and data systems that would feed these predictive machines and only use descriptive learning analytics in the meantime. Although there is still hope from some in the sector that predictive analytics could be developed in the future.

> to do proper predictive analytics, you need a big enough dataset…some of the work that we're definitely kicking off here in terms of improving our methods of ingestion of data from the source, being very picky about the quality of it. (EJ focus group)

This approach contrasts with what practitioners believed was happening in the commercial EdTech[1] sector (see e.g., Williamson, 2019) and with some internally developed solutions, where it was understood that demographic data were being used to fuel the predictions:

> [EdTech company] was using some of those protected characteristics. (EJ02)

> We've [a university with a self-built LA system] got a couple of different models for predictive learning analytics, one of which kind of predicts the likelihood of students completing their module. But it's based on, like, demographic data…I still do have some concerns, ethical concerns about labelling students. (EU2.01)

While protected characteristics data (e.g., gender, ethnicity, disability) was available, Jisc avoided using it because institutions and education practitioners expressed concerns 'about the ethics around' (Jisc P05) feeding predictive machines with such data. As practitioners at Jisc explained:

> A lot of institutions, they want to sort of almost ignore all that demographic stuff, which admittedly means losing sight of some quite important data, however ethically it seems like a better way to do it. (EJ Focus group)

For those institutions that perceived ethical issues with predicting student outcomes based on protected characteristics, it was believed by leading practitioners that they would need years to achieve good enough alternative data. This was because it involves ensuring that student records are populated with accurate data and regularly updated, cleaning datasets, improving access to systems, and feeding systems with additional sources of data. This makes it a challenging problem to address. Rather than pushing ahead with its predictive solution, Jisc has instead shifted its focus, believing it is better to help institutions to get their 'data right' (EJ01) before rushing to adopt predictive learning analytics solutions.

Addressing the 'data quality' issue

Some practitioners that we interviewed viewed this slower approach to adopting learning analytics in a positive way:

> We're doing it sort of slow and steady, to make sure that it fits for the institution. We're not just rushing into it, turning everything on and hoping for the best. We're doing it step by step, and making sure that it fits. (EU1.03)

This slow and steady approach, which goes against the grain of the acceleration logics (Rosa et al., 2017) surrounding much AI adoption, including that observed in the pharmaceuticals case, has included significant efforts to develop more accurate data about student engagement. This has led to the introduction of the time-consuming task of recording data about any interactions with students, something that could be frustrating, particularly in an already high-pressure working environment:

> For every student I have to log into JISC separately and say, I've sent a general email, I've sent a welcome back email… So it frustrates me, and I think a lot of my colleagues don't use it because of that… it's so much more work. (EU1.08)

Speculating on how this might play out in the future, practitioners at an institution already using predictive analytics observe that there was currently a problem with AI machines not capturing and integrating interventions based on the predictions:

> So, you have a model based on historical student behaviour, and on the basis of that you recommend an intervention… that will change that behaviour. And so instead of there being the predictive failure, there will be success…So you then feed that data back into the model for the following year and you've already started to undermine the model because …it triggered an action, [but] you haven't captured anything about what that action was, it's just that the data's changed. (EU2.03)

This same practitioner reflected on how this challenge might lead to future pressures on teaching staff to input more data into the machines:

> So, what that inevitably would mean is things like tutors have to record every phone call they make, every email they send, every reference to the student support team… in predetermined ways to be analysable. And so increasingly, rather than the tutor being concerned with the student, they're spending proportionally more and more time feeding data into a machine. (EU2.03)

Reflecting on this challenge in a focus group at the same institution, we see how these feeding the machine logics are becoming part of other educators' imaginaries of AI futures in their workplace:

> that is my concern, which is that we then start to change what it is we're asking the tutors to do, not because of the student, but to service the machine, if you like, to provide the data for the machine rather than directly supervising. (EU2 – focus group)

Unlike in the pharmaceutical sector, there did appear to be some flexibility for English HE institutions, supported by the non-profit Jisc, to cut a path away from the logic of acceleration of contemporary capitalism. This was evident in their decision to turn back from predictive analytics to take a slower and more considered approach. However, this meant that, as in the pharma sector, educational practitioners are increasingly being co-opted into efforts to feed the machine with high quality data, which they worried would be to the detriment of their everyday work with students.

Resisting extractivist logics in the arts

Use of AI in the arts

Similar to the previous two sectors, there is a long history of experimentation with machine learning and other AI technologies in the arts dating back to the 1960s and 1970s, with pioneers such as Vera Molnar and Harold Cohen (Broeckmann, 2019). However, interest in the use of AI in arts practice has escalated over the last twenty years, including most recently with the emergence of generative AI systems, particularly the text-to-image generators such as Midjourney and DALL-E (Weisz et al., 2023).

We observed that the data dilemma played out differently in the arts case relative to pharma and higher education. In many cases the artists we spoke with focused their practice on use of small and bespoke datasets, so had less concerns about data access to train their own models. The AI data dilemma identified was rather the one faced by firms in the tech industry that are developed GenAI products such as text-to-image generators and require significant amounts of data to train their models, including art works on which to train text-to-image generators. Their solution to this data access challenge has been indiscriminate harvesting of data from the internet without consideration of issues of ownership and labour to such an extent that many industry sources are now claiming this data source has largely been exhausted (see e.g., Posnett, 2025). Our focus in this section is on how arts practitioners who engage with various forms of AI in their work perceive this solution to the tech industry's data dilemma, and how their critique shapes their own AI practice.

The generative AI pressure point in the arts

Despite excitement about experimenting with GenAI tools from some practitioners (A13), most of our participants strongly critiqued the feeding the machine logic involved in the development of GenAI tools such as text-to-image generators. These concerns were particularly focused on the extractive dynamics of new generative AI models (i.e., powerful tech firms practice of harvesting and using vast quantities of data to train models without consideration of issues such as labour, compensation, consent, see e.g., Crawford, 2022), but also extended to AI techniques in general:

> I think definitely the authorship problem needs to be addressed. This kind of rampant data scraping and stealing of people's life's work is not going in the right direction [laughs]… there's no legal repercussions for scraping data. (A04)

They highlighted the lack of legal repercussions for scraping data with which to feed the machine as a significant problem in relation to ethical concerns and values such as accountability, fairness, responsibility and transparency (A02, A03, A04, A05, A06, A07, A08, A10, A12, A14). Many discussed their perception that authorship and intellectual property land on artists differently, compared to practitioners in other sectors:

> With text to image tools it's so easy to rip off particular styles of certain artists or ways of doing things, literally just like putting their name in the prompt…I think there's a whole generation of people …that just don't seem to think that there's any problem with doing this, or don't really see why that would, you know, rub people up the wrong way who had been spending years or decades developing a certain kind of practice or style. (A10)

Values driven critique and practice as a response to extractivism

Despite working in a context impacted by the emergence of big tech's GenAI (Michaels, 2024), the majority of participants were not engaged in commercial work, so avoided many of the corporate and regulatory pressures evident in other cases in their day-to-day work. This allowed them greater freedom to express critical viewpoints and to experiment with how they used AI in their work. Critiquing capitalist assumptions and standing in opposition to extractivist logics of Big Tech, arts practitioners primarily focused on the societal values and implications of AI in the arts and wider context:

> I would say I'm like 30% optimistic and 66–33% optimistic and 66% pessimistic. And that's because, you know, society

> is dominated by capital and, you know, things are done for profit, not the good of society. (A02)

These more critical perspectives enabled participants to acknowledge their role and place within the capitalist system and be reflexive about what this meant for their artistic practice. Some artists resultantly focused their practice solely on systems that diverged from hegemonic and capitalist structures, such as using open source tools and their own curated datasets as a means to work against big tech monopolisation. On the other hand, while profit driven values such as optimisation were not mentioned by any of our participants, similar to practitioners in the pharmaceutical case many of them felt pressured to keep up with the rapidly developing technology and field:

> The field is changing week by week, with new models, higher fidelity possibilities and more real time processing options. Keeping abreast of can be a challenge and knowing where to look for papers, new models and centres of learning is important. (A08)

Despite these frustrations, many practitioners tried to develop their practice by navigating the pressure of the capitalist logics by approaching feeding the machine differently, informed by values geared towards making a positive contribution to society. Most opted for small, careful, ethical, and often their own data, rather than relying on big datasets, faster throughout and large-scale appropriation of others' work:

> The idea of using massive resources to process data in order to make financial decisions or, like, weird trippy artworks, I think that's not really going to be the answer to our problems …So I don't think that the kind of race for faster and faster and more throughput is really the long-term future for machine learning in the arts. I think probably small data approaches. (A01)

As part of the ethical values guiding participants, environmental concerns and consequences (Brevini, 2021) were also significant. These concerns included carbon dioxide emissions resulting from the extensive energy consumption of data collection and AI training models – another effect of feeding the machine:

> The growth of data centres is wild, and there's not really any pushback against those developments. A lot of them are using coal power. A lot of them just directly tap into the kind of extractivist industry of mining around the world, whether for energy or for minerals, for computer manufacturing… [it] has just made me think about…what kind of strategies or ways of combating that can be used or in place. (A14)

While the arts practitioners we spoke with were explicit in trying to navigate feeding the machine in line with their values, this proved challenging given their awareness of their entanglement in the capitalistic system and extractivist industry. Their concerns about the ethical use of data and AI lacked straightforward solutions, as highlighted by many arts practitioners. Moreover, a more commercially orientated artist, who was at risk of losing his job due to new generative tools, expressed dissatisfaction with the narrative of 'small data' as a form resistance:

> In terms of curation…there'll be essentially crap like this, where it says, 'So and so used a curated, handcrafted dataset to explore this or that'…the discursive blurb often tries to frame it as this kind of couture-like version of what's happening in industry, thus legitimising it and giving it value. I would debate this. I would say that actually this is an example of contemporary art bullshit. (FG 2)

While participants in the arts sector were situated differently within the creative industry, they all, to varying degrees, tried to reconcile the demands of the capitalist model with the creative potential of AI by adhering to their personal values. The willingness to use and experiment with AI tools, alongside a critical approach to AI technologies, led many of our participants to express mixed and often contradictory emotions when asked about their artistic practice:

> I think probably my feelings are fairly clear, but if I had to state them, I feel uncertain. I feel anxious. I feel existential. I feel a small amount of excitement. I feel exhausted. I feel threatened. I feel curious, resentful. Yeah. (A13)

Similar to the pharmaceutical and higher education sectors, ethical engagement with data is crucial in the arts. However, unlike other sectors, arts practitioners strongly experience the pressure of the extractivist-generative AI moment that is shaped by big tech's approach to resolving its own AI data dilemma. Moreover, they possess distinct flexibility in feeding the machine with what they believe to be ethically appropriate data, achieving this by focusing on small-scale datasets, often which they have produced themselves.

Discussion and conclusion: The costs of feeding the machine

While the tech industry, media, and government promote a future transformed by AI, concerns are mounting regarding 'AI readiness' in real-world contexts (e.g., McKendrick, 2024; Turing Institute, 2024). Building upon existing critical scholarship about practitioners' experiences of and desires about the integration of AI and other algorithmic systems into workflows, this paper addressed a specific aspect of this readiness challenge: the 'data dilemma'. That is, the lack of sufficient appropriate and ethically sourced data required to train AI algorithms effectively; a challenge

which undermines the feasibility of widespread AI adoption across sectors. We identified examples of this dilemma in three contrasting contexts: pharmaceuticals, higher education, and the arts. Each case offers a unique perspective on how the 'data dilemma' can manifest and how existing data practices within each sector are not always suitable to feed the AI machine. Our findings bring to the fore a further AI challenge beyond existing concerns about e.g., exploitation of precarious labour (e.g., Graham, 2017; Gray and Suri, 2019), algorithmic discrimination (e.g., Buolamwini and Gebru, 2018; Noble, 2018; Barocas and Selbst, 2016), and environmental harm (e.g., Brevini, 2021).

We argue that, at its core, the data dilemma presents as a lack of appropriate data for viable, ethical and desirable AI integration. In parts of the pharmaceutical industry, this translates to a scarcity of data about 'bad' compounds. Educational institutions struggle with capturing a rich and accurate picture of student and staff activity, without resorting to predictions based on demographic characteristics. The artistic domain faces a somewhat different challenge, that of generative AI firms harvesting online artworks to overcome their data dilemma and bolster training datasets, and in doing so continuing a longstanding dynamic of capitalist appropriation (see e.g., Rosa et al., 2017) that is familiar to creators, both in the context of the internet (e.g., see controversies around Google Books, Spotify) and beyond.

In the examples from the three sectors, we saw labour being harnessed to overcome this dilemma, whether directly from employees, or indirectly through data extraction. We also found that this drive often overlooks the human cost of overcoming these data limitations and the shorter-term productivity implications as highly skilled practitioners are drawn into data production tasks such as creating detailed student records and running 'soul destroying' failing experiments, at the expense of activities that practitioners find more meaningful. We observe this in the profit-driven corporate context of the pharmaceutical industry and the neoliberal regulatory environment of higher education, as well as in the new forms of extractive practices taking root in the arts. In each of these contexts, practitioners experience different forms of pressure to fill the data gaps required by AI-driven capitalism, either actively (i.e., pressure to refocus attention on new data production practices) or passively (i.e., pressure to accept the data harvesting practices of firms whose values they oppose). A further issue to be explored in future research is to understand better who becomes responsible for these burdens. It is well established that marginalised employees often experience a higher burden of unrewarding tasks and appropriation in organisational contexts, and it is necessary to consider to what extent this pattern is playing out in cases where practitioners are put to work feeding the machine.

Our research also highlights the emergence of different modes of resistance against these growing pressures to feed the machine. While the pharmaceutical industry exhibited minimal resistance beyond raising concerns with colleagues, likely due to corporate constraints, in the HE sector we see institutions and practitioners resisting feeding learning analytics machines with demographic data given the ethical risks of bias and discrimination. We also see suggestions that there was not full compliance among practitioners for some data entry practices being promoted within the HE sector. In the artistic domain we saw practitioners advocating for 'small data' and open-source approaches for artists experimenting with AI, as an alternative way of feeding the machine – doing AI differently to GenAI firms. These dynamics of resistance also reflect how the pressure on practitioners to actively feed the machine was stronger the more closely the sector was embedded within neoliberal capitalist value systems of, for example, profit, acceleration, efficiency and compliance. We found pressure and lack of agency most evident in the pharmaceutical sector and least evident in the arts where independent practitioners had more control over their own practices despite little say over the activities of extractivist big tech.

These findings contribute to a deeper understanding of AI's 'data dilemma', as well as to broader research agendas examining the implications of AI adoption for practitioners, organisations' desire for numbers and the nature of data inputs discussed in the early sections of this paper. Building on this foundational CDS literature, we examined the human implications of the AI data dilemma in everyday practice. These implications can be added to the growing list of challenges for AI adoption, drawing attention to the ways that AI's data dilemma risks drawing highly skilled practitioners into activity that prioritises feeding the machine over the more humanistic, creative and rewarding elements of their labour. In resource-constrained environments, expecting experienced practitioners to shoulder this burden of data creation to overcome the data dilemma can lead to demoralisation and ultimately, losses in productivity. Whereas in the arts sector, unconstrained harvesting of art works from the internet risks the viability of careers within the commercial arts such as illustration. While clearly there are positive and hopeful moments in our stories of AI integration, such losses risk the demise of what is of value in the skilled labour of practitioners (e.g., creativity, connection, deep thinking), in exchange for the possibility of future profits and regulatory compliance that our current political economic system demands.

The question then becomes what are the alternatives? While one approach practitioners may adopt is to wait out the hype cycle, focus on the positives and hope the pressure abates, this would amount to a form of digital resignation (Draper and Turrow, 2019) with practitioners abandoning any agency they might have to influence how AI integration plays out in their contexts of practice. Instead, we suggest the possibility of building relations of solidarity with others experiencing any of the various negative impacts of AI integration on their everyday lives and working practices, so as to explore and develop alternative ways of imagining and doing AI. The artists we spoke to are already beginning to do that work of imagining alternatives, as are many other

groups (e.g., Jones and melo, 2021; Carceral Tech, n.d). Such efforts can promote the development of AI in directions that align better with some of the alternative value systems that practitioners appear to appreciate (e.g., values such as autonomy, creativity, ethics, slowness, carefulness) and lay down lines for a politics of refusal of practices and technologies that are not in the best interests of people and planet. While this is undoubtedly the more challenging approach it is the one that may eventually lead to a more just, sustainable and humanistic consideration of when and how to integrate AI technologies into society.

ORCID iDs

Jo Bates https://orcid.org/0000-0001-7266-8470
Monika Fratczak https://orcid.org/0000-0001-6156-0982
Helen Kennedy https://orcid.org/0000-0003-0273-3825
Itzelle Medina Perea https://orcid.org/0000-0003-1702-1484
Erinma Ochu https://orcid.org/0000-0002-7268-278X

Ethical statement

The research was ethically approved by the University of Sheffield.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported in the paper was funded by the Arts and Humanities Research Council, UK (grant number: AH/T013362/1).

Declaration of conflicting interests

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

Note

1. Education Technology industry which develops technological tools and resources aimed at enhancing learning and teaching, one example of which is learning analytics platforms which are used in Higher Education.

References

Barocas S and Selbst AD (2016) Big data's disparate impact. *California Law Review* 104(671): 671–737.

Bates J, Lin YW and Goodale P (2016) Data journeys: Capturing the socio-material constitution of data objects and flows. *Big Data & Society* 3(2).

Benjamin R (2019) *Race After Technology: Abolitionist Tools for the New Jim Code*. Cambridge: Polity Press.

Bowker G (2008) *Memory Practices in the Science*. Cambridge, MA: MIT Press.

Braun V and Clarke V (2006) Using thematic analysis in psychology. *Qualitative Research in Psychology* 3(2): 77–101.

Brevini B (2021) *Is AI Good for the Planet?* Cambridge: Polity Press.

Broeckmann A (2019) The machine as artist as myth. *Arts* 8(1): 25.

Buolamwini J and Gebru T (2018) Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 81: 77–91.

Carceral Tech (n.d.) Carceral Tech Resistance Network. Available at: https://www.carceral.tech/ (accessed 23 September 24).

Chan J, Sanders C, Bennett Moses L, et al. (2022) Datafication and the practice of intelligence production. *Big Data & Society* 9(1).

Choroszewicz M (2022) Emotional labour in the collaborative data practices of repurposing healthcare data and building data technologies. *Big Data & Society* 9(1).

Crawford K (2022) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press.

Currie ME, Paris BS and Donovan JM (2019) What difference do data make? Data management and social change. *Online Information Review* 43(6): 971–985.

Dencik L and Stevens S (2021) Regimes of justification in the datafied workplace: The case of hiring. *New Media & Society* 25(12): 3657–3675.

Draper NA and Turow J (2019) The corporate cultivation of digital resignation. *New Media & Society* 21(8): 1824–1839.

Fine GA (2007) *Authors of the Storm: Meteorologists and the Culture of Prediction*. Chicago: University of Chicago Press.

Gabrys J, Pritchard H and Barratt B (2016) Just good enough data: Figuring data citizenships through air pollution sensing and data stories. *Big Data & Society* 3(2).

Gitelman L (2013) *Raw data is an oxymoron*. Cambridge, MA: MIT Press.

Gov.uk. (2023) Sponsorship duties (accessible). Available at https://www.gov.uk/government/publications/student-sponsor-guidance/sponsorship-duties-accessible (accessed 23 September 24).

Graham M, Hjorth I and Lehdonvirta V (2017) Digital labour and development: Impacts of global digital labour platforms and the gig economy on worker livelihoods. *Transfer: European Review of Labour and Research* 23(2): 135–162.

Gray ML and Suri S (2019) *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt.

Heymann M, Gramelsberger G and Mahony M (2017) *Cultures of Prediction in Atmospheric and Climate Science: Epistemic and Cultural Shifts in Computer-Based Modelling and Simulation*. London: Taylor & Francis.

Hoeyer K (2019) Data as promise: Reconfiguring Danish public health through personalized medicine. *Social Studies of Science* 49(4): 531–555.

Jarke J and Büchner S (2024) Who cares about data? Data care arrangements in everyday organisational practice, information. *Communication & Society* 27(4): 702–718.

Jones ST and Melo NA (2021) We tell these stories to survive: Towards abolition in computer science education. *Canadian Journal of Science, Mathematics and Technology Education* 21: 290–308.

Kennedy H (2016) *Post, Mine, Repeat: Social Media Data Mining Becomes Ordinary*. London: Springer.

Kennedy H and Hill RL (2018) The feeling of numbers: Emotions in everyday engagements with data and their visualisation. *Sociology* 52(4): 830–848.

Kitchin R (2014) *Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

López Jiménez EA and Ouariachi T (2021) An exploration of the impact of artificial intelligence (AI) and automation for communication professionals. *Journal of Information, Communication and Ethics in Society* 19(2): 249–267.

Lupton D (2019) It's made me a lot more aware': A new materialist analysis of health self-tracking. *Media International Australia* 171(1): 66–79.

Mackenzie A (2015) The production of prediction: What does machine learning want? *European Journal of Cultural Studies* 18(4-5): 429–445.

McKendrick J (2024) Data is the missing piece of the AI puzzle. Here's how to fill the gap, ZDNet. Available from: https://www.zdnet.com/article/data-is-the-missing-piece-of-the-ai-puzzle-heres-how-to-start-filling-the-gap/ (Accessed 23 September 2024).

Medina Perea I (2021) Socio-material factors shaping patient data journeys in the United Kingdom. PhD thesis, University of Sheffield. https://etheses.whiterose.ac.uk/30111/.

Michaels C (2024) The art world's AI dilemma: how can artists and museums thrive when big tech controls the monetising of artificial intelligence? The Art Newspaper [online] 30 May. Available at: https://www.theartnewspaper.com/2024/05/30/the-art-worlds-ai-dilemma-how-can-artists-and-museums-thrive-when-big-tech-controls-the-monetising-of-artificial-intelligence (accessed 13 August 2024).

Muldoon J, Graham M and Cant C (2024) *Feeding the Machine: The Hidden Human Labour Powering AI*. Edinburgh: Canongate Books.

Noble S (2018) *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: NYU Press.

Office for Students (2024a) Data and analysis. Available at: https://www.officeforstudents.org.uk/data-and-analysis/. Accessed 23/9/24.

Office for Students (2024b) Navigating financial challenges in higher education. https://www.officeforstudents.org.uk/publications/navigating-financial-challenges-in-higher-education/ (accessed 13 August 2024).

Pantzar M and Ruckenstein M (2014) The heart of everyday analytics: Emotional, material and practical extensions in self-tracking market. *Consumption Markets & Culture* 18(1): 92–109.

Passi S and Sengers P (2020) Making data science systems work. *Big Data & Society* 7(2).

Pine K, Bossen C, Holten Møller N, et al. (2022) Investigating data work across domains: New perspectives on the work of creating data. *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (CHI EA '22)* Article 87: 1–6.

Posnett K (2025, May 19) The new markets for AI data. Financial Times. Available at: https://www.ft.com/content/625b0a98-a68d-49b6-b063-2179e3cb77f0.

Roberts ST (2019) *Behind the Screen*. New Haven: Yale University Press.

Rosa H, Dörre K and Lessenich S (2017) Appropriation, activation and acceleration: The escalatory logics of capitalist modernity and the crises of dynamic stabilization. *Theory, Culture & Society* 34(1): 53–73.

Stalph F (2020) Evolving data teams: Tensions between organisational structure and professional subculture. *Big Data & Society* 7(1).

Thylstrup NB (2022) The ethics and politics of data sets in the age of machine learning: deleting traces and encountering remains, Media, Culture andamp. *Media, Culture & Society* 44(4): 655–671.

Turing Institute (2024) Seminar Turing Institute: AI's Data Dilemma: Confronting the Paradox of Poor Quality Data in the Age of AI. Fri, 26 Jan 2024. Available at: https://www.turing.ac.uk/events/ais-data-dilemma-confronting-paradox-poor-quality-data-age-ai.

Turja T, Minkkinen J and Mauno S (2022) Robotizing meaningful work. *Journal of Information, Communication and Ethics in Society* 20(2): 177–192.

van Dijk J (2013) *The Culture of Connectivity: A Critical History of Social Media*. Oxford: Oxford University Press.

Weisz JD, Muller M, He J and Houde S (2023) *Toward General Design Principles for Generative AI Applications*. Available at: https://arxiv.org/pdf/2301.05578.pdf.

Williamson B (2019) Policy networks, performance metrics, and platform markets: Charting the expanding data infrastructure of higher education. *British Journal of Educational Technology* 50(6): 2794–2809.