Edmond, Jennifer, Nicola Horsley, Jörg Lehmann, and Mike Priddy. "From obscure data to datafied obscurity: The invisibilities of datafication." *The Trouble With Big Data: How Datafication Displaces Cultural Practices*. London,: Bloomsbury Academic, 2022. 105–133. *Bloomsbury Collections*. Web. 26 Apr. 2023. <<http://dx.doi.org/10.5040/9781350239654.ch-6>>.

Downloaded from Bloomsbury Collections, www.bloomsburycollections.com, 26 April 2023, 14:38 UTC.

Copyright © Jennifer Edmond, Nicola Horsley, Jörg Lehmann and Mike Priddy 2022. Released under a CC BY-NC-ND licence (<https://creativecommons.org/licenses/by-nc-nd/3.0/>). You may share this work for non-commercial purposes only, provided you give attribution to the copyright holder and the publisher, and provide a link to the Creative Commons licence.

From obscure data to datafied obscurity: The invisibilities of datafication

We have discussed in the foregoing chapters how the promise of the data ‘deluge’ is often divorced from its dynamic relationship with the humans who are both (active or passive) producers and users of data and its conduits into wider cultural practices. The dominant messages surrounding big data often neglect the complexities of people in the process – from HeLa’s disembodied *immortal cell line* to the ‘artificial artificial intelligence’ of Amazon’s Mechanical Turk – eliding human contributions to position the outputs of datafication (portrayed as a biospheric black box like a pure Internet of Things) centre stage. This chapter looks at where applying the principles of data science to cultural heritage may take our relationship with complex knowledge. A comprehensive view of the impacts of datafication – both on the archival practices that sustain the historical record as a resource, and on the practices adopted by users of that resource – reveals cultural shifts that risk closing us off from complexity. We examine how the inductive imperative of big data applied in the sphere of business and elsewhere crosses over to the cultural realm as an example of how deference to human reasoning is being displaced, to the extent that the use of AI to determine what we learn from the historical record might soon be commonplace, and why exactly that is so troublesome.

As we saw in Chapter 2, data is often represented as clinically controlled and removed from the awkward complexities of human actions and interactions, especially when it is *of* those things and used to understand them. The archivists and related practitioners who work with cultural heritage data – on whom Cook and Schwartz observed a pressure to be a ‘white-coated clinician’ they saw as inherent in a focus on the ‘mechanics’ of their data practices¹ – exemplify the role of accommodating complex knowledge in the data ecosystem. This position makes them an untapped resource in the project of understanding how big data is changing our world, for good or for ill.

Recognizing archivists as actors who influence the space in which they operate is at least as vital as acknowledging the positionality of the researchers who use their services. Furthermore, it would be a mistake to judge practitioners' turn towards applying certain tools to their 'inherently chaotic'² world as a failed enterprise because of the persistence of knowledge complexity. As Joks, Ostmo and Law,³ drawing on Latour,⁴ describe with vivid imagery of the complexities of *simply* counting fish (simpler, surely, than shooting them in a barrel), even in 'hard' science, 'practices are fluid and hard to predict ... scientific practices are art forms too [yet] the fluidities embedded in scientific practice vanish ... fluidity disappears once inscriptions get reified ... people start talking about science as if it were stable all the way down'.⁵ Such stable 'objectivity' has been much critiqued. Donna Haraway, for example, argued against the promise of 'vision from everywhere and nowhere equally and fully, common myths in rhetorics surrounding Science', which she found to be characteristic of a 'god trick' approach to truth as either total or relative.⁶ Haraway instead advocated a 'situated' approach that acknowledges the 'partial perspective', wherein:

science becomes the myth, not of what escapes human agency and responsibility in a realm above the fray, but, rather, of accountability and responsibility for translations and solidarities linking the cacophonous visions and visionary voices that characterize the knowledges of the subjugated.⁷

As the norms of datafication pervade our approaches to every field of knowledge, the translations, solidarities, visions and voices that occur, receive validation or become subjugated in the cultural data environment allow us to comprehend the tacit tensions that reveal the depth and scale of big data's impact on our culture and our ability to envision what 'might have been otherwise'.⁸

What you see is what you get

The hypervisibility that the current era lends certain data (such as that 'freely available' on social media platforms⁹) has been subject to critique from many quarters. As we have described in Chapter 3, the hyperabundance of data that is being made accessible is not easily navigated but necessitates the processes and tools of sensemaking. In addition, the big data era presents a new landscape where entrenched values, biases and assumptions are hard-wired and coalesce to cast shadows over some of the resources of our collective memory. In Chapter 2, we discussed how crucial metaphors are in reproducing

or disrupting our culture. When considering the importance of analogies to AI, Pearl asserts that, despite the admirable accuracy of Babylonian ‘black-box predictions’ in advancing astronomy, it was the creative power of the ancient Greeks’ metaphors that allowed them to surpass such a ‘data-fitter’ approach in order to ‘model’ the calculation of the Earth’s circumference.¹⁰ Bowker explains the power of the metaphor thus:

Our knowledge professionals see selfish genes because that’s the way that we look at ourselves as social beings – if the same amount of energy had been applied to the universality of parasitism/symbiosis as has been applied to rampant individualistic analysis, we would see the natural and social worlds very differently. However, scientists tend to get inspired by and garner funding for concepts that sit “naturally” with our views of ourselves.¹¹

When we apply principles of human thought to new technologies then, as in the definition of AI as ‘computing technologies that resemble processes associated with human intelligence’,¹² we imbue those technologies with such values. The advent of an era of ‘scientific discovery by computer’ that ‘is totally isolated from all social and cultural factors whatever’ was heralded by Slezak in 1989, who posited that AI constituted ‘a “pure” or socially uncontaminated instance of inductive inference’.¹³ The march of the inductive imperative across academic disciplines has been observed by commentators such as Kitchin, who describes the creep of a ‘new empiricism’ driven by ‘agnostic data analytics’ that allow data to ‘speak for themselves free of human bias or framing’ – the results exposing a pure truth.¹⁴ Kitchin sees this paradigm shift extending all the way across the social sciences and the humanities,¹⁵ areas we would expect to be impervious to the ‘myth of big data’: the sirens’ call promising to free us from human bias.¹⁶ As we discussed in Chapter 3, bias is central to historical enquiry, the historian’s navigation of subjectivities demonstrating the verifiability of her methods through the ‘persistent plodding’¹⁷ of her own decision making. We have identified concerns associated with the separation of understanding from knowledge when the black-boxed decision making of neural nets’ predictive approaches are used – an extra layer of opacity from stochastic data models.

Concerns have been raised about big data tools since long before the application of neural nets reached deep learning’s current level of sophistication. In 1993, Neil Postman drew on Frederick W. Taylor’s work to describe the principles of Technopoly, which revolved around the primacy of efficiency as the goal of human physical endeavour and thought, and prized measurement and machine calculation over human judgement, which was seen as flawed and

unnecessarily complex. Postman was concerned about the growth of the ideal conditions for Technopoly to take hold, namely the appearance of data divorced from their context and intended audience, travelling in no particular direction 'in enormous volume and at high speeds, and disconnected from theory, meaning, or purpose'.¹⁸

As we saw in Chapter 3, data used by scientists is conventionally collected with the purpose of answering a specific research question, whereas humanists are often inventive in discovering the buried treasure of knowledge in sources that were ostensibly created and preserved for other reasons. The big data era arrived with an acknowledgement of the abundance of data creation that was now humanity's habit and the hard-to-dispute suggestion that perhaps this might be a massive resource if properly (scientifically) managed and used to investigate questions with which it might not otherwise be associated (in the tradition of humanist hermeneutical methods). Researchers were urged to consider reusing the data of existing sources before designing new studies,¹⁹ and the reassuring sanctity of the scientific method in ensuring against incorrect data use, coupled with threats to qualitative research including an increasingly competitive funding environment and 'survey fatigue' among over-researched populations, was persuasive. Archival institutions' moves to accommodate the digital turn demonstrate how strongly societal-level knowledge trends cross into more specialized academic research domains (and how fruitless it is to consider the flows of common knowledge practices in isolation from each other) with the ubiquity of Google-driven search habits the most compelling example, as we saw in Chapter 4.

The standardization of interfaces between knowledge-seekers and myriad knowledge institutions has obfuscated huge differences in the organization, values and practices of those institutions. Many elements of archival practitioners' work go, unsurprisingly, unsung, but the drive to furnish users with detailed information about collections without having to ask for it suggests that dialogic exchange with these gatekeepers was an unnecessary barrier that has been removed. There is a fundamental shift here in access to knowledge and knowledge creation power structures: users begin to feel closer to the knowledge they seek while practitioners drift away from their pivotal position as their grasp of how material is being used loosens. If this sounds like a democratising shift (and we will discuss this later), any such benefit is hobbled by the reality that an overwhelming amount of data that does not become formally recorded as metadata continues to be stored as tacit knowledge in cultural heritage practitioners themselves. The continuing daily practices of responding to

user enquiries at institutions with the most developed online catalogues are testament to this tacit knowledge. As well as exposing how the different systems behind standardized interfaces vary in completeness and accuracy, this human labour is a product of the specific complexities individual knowledge institutions specialize in and the ways in which these complexities inform their organizational cultures. In this sense, the continuing reliance on tacit knowledge in unearthing hard-to-find material is not simply a matter of jealously guarded professional identity, which leads to losses as knowledge either is left behind when it does not fit the mechanisms of routinization or travels with the person when they leave an institution (it is also noteworthy that, across the cultural heritage sector, many essential tasks that sit between the routinizable and the professional specialist are performed by volunteers).

The changing use of collections is developing in tandem with new skills and vocabularies through which imaginaries of knowledge might be articulated. This was seen to be deconstructing the research process as data archives as well as conventional cultural heritage institutions incorporate data repositories which, alongside Figshare, Zenodo and others, and with the help of research infrastructures, promote the sharing of data from research studies as separate from the researchers' prestige publications. Knowledge creators' use of these opportunities to share data is creeping up across academic fields but the competencies required for their discovery remain unevenly spread. In large cultural institutions such as libraries whose holdings number many millions of items, understanding the relationships within and across collections can be challenging for the uninitiated. Within the KPLEX interviews, one head of a service team at a national library described how digitization presented the mammoth challenge of impressing the importance of context upon a user who has *landed* on a page without an understanding of how what they are viewing relates to the institution's collections as a whole (never mind the collections of other institutions). It was felt that the library's traditional visitors forged an awareness of the number of *boxes of stuff* that related to their query versus the number they had actually got to grips with, whereas today's user, presented with a satisfying Google lookalike result, has her curiosity curtailed.

Furthermore, independent research experimenting with language technology like Open AI's GPT-3²⁰ has suggested that a tendency in the popular imagination to value the immediacy of 'solutions' over accuracy in terms of how germane they are to the spirit of a query, coupled with black-boxed mechanisms, creates the conditions for powerful AI to steer towards goals that are neither optimal nor apocalyptic, but skewed by placing too much value on 'quick wins'. This was

described on vox.com as ‘handing over our future’ by mistake, ‘one it’d be easy to make step by step, with each step half an accident’.²¹

When working with data engineers to change systems and practice at diverse knowledge institutions, cultural heritage practitioners are keenly aware of the challenges outlined above. Digitization is a much more involved process than the mere production of a digital copy of artefacts and a careful balance must be struck to ensure that, in preparing material to be read by a machine, knowledge complexity is not omitted or dodged in favour of producing a manageable simulacrum that does not do justice to the item’s links to – or divergence from – other knowledge, or significant milestones on its journey to the user. Once (some of) an institution’s collections are digitized, it can join a data-sharing infrastructure that will take the heavy lifting out of broadcasting its new digital discoverability to the world. These initiatives peel back another layer of context by applying further standardization. Kitchin describes databases and digital infrastructures as ‘unmooring’ analysis from an understanding of data production and representation, with standardization of metadata for processing by algorithms working to ‘decontextualize and depoliticize the data contained within’.²² For practitioners, opening access can result in a weakened relationship between the institution and its users, raising questions about gatekeeping and control over more or less porous boundaries as the historical record becomes untethered from a physical location. Opening access to the world’s knowledge is, however, the *raison d’être* of cultural heritage institutions. The following sections explore how the tension between a commitment to expanding knowledge and an awareness of the risks of doing so through current technologies manifests in two forms of invisibility.

The minoritized material: Corner cases and downward spirals of invisibility

The imperative for digital discoverability and growth of data infrastructures has led to the sharing of cultural knowledge sources from institutions of all sizes and specialisms, raising the profile of sources of knowledge that had previously been limited to an intellectual or geographic niche. Cultural institutions themselves may experience a boost in their visibility, especially smaller museums, libraries and archives accustomed to serving a local or site-specific audience, but their new profile may not tell the whole story. Digitization is a costly process in terms of budget, time and labour intensiveness, which includes the curatorial

practitioners, who are the experts on the collections of a museum, library, gallery or standalone archive, working with data engineers to build systems that reorganize and reconstitute holdings and metadata to facilitate digital sensemaking techniques – with the burden usually on archival practitioners’ *upskilling* to understand computational thinking.²³ Despite high levels of cooperation and apparent agreement on goals, institutions often share only a fraction of their collections.

Practical considerations drive decisions about which of an institution’s holdings will be digitized *first*, and which are left on the backburner, perhaps for the next phase of a project, in the event of further funding or – more or less intentionally – never to see the light of digitization. Budget constraints are a very real feature of the cultural sector (although experienced very differently across countries, regions and specialisms) and priorities must be set and justified. In making these decisions, however, it is clear that knowledge practitioners’ perceptions of the possibilities, risks and threats of big data technologies, up to and including AI, come into play.

In moving away from *the material*, practitioners fear knowledge complexity may be under-exposed and skimmed over. A Matthew effect²⁴ might then stimulate repeated use of data that is more readily discoverable in a fashion similar to an article that achieves a high position in PageRank maintaining its popularity through this exposure. Practitioners’ acute awareness of the kinds of nuance inevitably lost in the distillation of knowledge into data leads them to doubt that the special qualities certain items have to offer the casual knowledge seeker will be picked up on by the mediating technology. The threat here is that some forms of knowledge will not be properly translated by computational methods and, although rendered technically discoverable, the features that make them an attractive, perhaps unique, source will not end up as metadata that provide a clear path for users who do not already know what they are looking for. Practitioners’ expertise moves them to further caution through their imaginaries of a machine that proffers some artefacts *too* readily, that is, one that enables undesirable data linking. The dark side of discoverability is characterized by practitioners as the potential for ‘artificial intelligence ... to draw new conclusions’, particularly from unstructured data, which had, until recently, ‘resisted the broader analysis’. With this comes the ‘significant danger’²⁵ that the inferences of machine processing and data linking practices that are the data science modus operandi of Google and Facebook could expose identifying data about individuals. Digital discoverability therefore exposes a dark side that archivists are used to mediating as gatekeepers of material that is vulnerable to

misuse and magnifies this threat beyond the potential that can be seen unaided by digital tools. Unsurprisingly, practitioners have acted on this understanding to resist certain moves towards knowledge sharing. Self-styled ‘data skeptic’, Cathy O’Neil, explains this mentality in her message to ‘the overly sunny data lover’ – a reminder that not everything we seek to understand is

measurable, that not all proxies are reasonable, and that some models have unintended and negative consequences. While it’s often true that doing something is better than doing nothing, it’s also dangerously easy to assume you’ve got the perfect answer when at best you have a noisy approximation.²⁶

Unintended consequences in general, and *unreasonable* proxies and *imperfect*, satisficing answers in particular, are at the heart of cultural knowledge practitioners’ reservations about datafication beyond the dark side of discoverability. Underlying these concerns is a perception familiar to readers of Latour and Callon’s work *How to Follow Scientists and Engineers*, which observed how differences in practice were translated into technical problems that engineers could then apply technological ‘solutions’ to – a phenomenon sometimes referred to as ‘techno-solutionism’²⁷ (see also Layne’s work on the ‘cultural fix’²⁸).

Simply adapting technologies that have successfully worked with certain data for use with artefacts from galleries, libraries, archives and museums minimizes the likelihood of adopting the optimal approach for dealing with unique historical holdings. Even where the choice of technology is not necessarily inappropriate, AI’s seemingly unassailable march across data frontiers has stumbled due to inadequate training. Biases towards assumed audiences and purposes of services lead to user problems such as those experienced by people with Scottish, Indian and even some US accents when using Apple’s personal assistant app. Siri was tested in-house, its English language recognition progressing iteratively from the *default* American accents of Apple employees.²⁹ It is therefore a reasonable expectation that data at the fringes of developers’ familiarity will not be *understood* by a machine working towards these foreign poles from the smooth terrain at the centre of its given map.

As we will discuss further in the next chapter, minoritized languages are at risk of suffering from a Matthew effect due to their deviance from the default, and it is understandable that experts in fields whose lexicons use specific terms in a different way from their use in the common vernacular may be alert to the danger, not just of marginalization but of erroneous linking to other meanings by a learner AI that misconstrues such specificities. Language use is a key

vulnerability of design for narrative knowledge but the stories of cultural heritage are told through many formats that have proved challenging for algorithms. In 2015, Google announced it was ‘appalled and genuinely sorry’ after its Photos service labelled images of black people as featuring ‘gorillas.’ More than two years later, a Wired investigation found that Google’s ‘fix’ had simply erased gorillas, chimpanzees and monkeys from the Newspeak vocabulary the image recognition algorithm was allowed to draw on.³⁰ As an example of the appetite an extremely well-resourced leader in the tech field has to properly serve ‘corner cases’ that problematize the supposed sophistication of its AI modelling, the case of people versus primates hardly inspires those asked to put their trust in such black boxes. This dilemma goes beyond corner cases to even more insidious biases in the associations AI is trained to make, as demonstrated by Google Vision Cloud’s labelling of a hard-to-identify object as a ‘monocular’ when in the hands of a ‘white’ person (with 60 per cent certainty of this – correct – recognition) but more likely a ‘gun’ in the hands of a black person (with 61 per cent probability).³¹ Similarly, an experiment by the American Civil Liberties Union³² found that Amazon’s Rekognition facial recognition system incorrectly matched twenty-eight members of Congress to mugshots from a police database, disproportionately misidentifying those who were not white.³³ These cases’ clear implications of racial discrimination (and persistent work by pioneering researchers like Joy Boalamwini and Timnit Gebru³⁴) raised the spectre of datafication as obscuring the representation of minorities but other research has shown how this trouble with big data is fundamentally *cultural*: Beery et al., for example, found recognition algorithms to have no problem identifying a cow when it grazed on an Alpine pasture but apparently clueless when one appeared on a beach (see Figure 6).³⁵

Of course, casting the fresh eye of AI over images whose use has been delimited by the descriptive text metadata assigned to it at the time and place of their assumption into the historical record offers the possibilities of new uses as well as risks. A recent example is Google’s decision to assume a gender-blind approach to images that is potentially liberating for their subjects.³⁶ While there may have been good reason for a human describer to omit certain details in a given context, the possibility for recognition of new connections is key to the hopeful ethos of the big data revolution. Custodians of cultural knowledge are used to aiding revivalists while combatting revisionists and although repositioning users closer to the data and further away from gatekeepers might increase vulnerabilities to misuse (for example, holocaust denial), the founding principles of memory institutions – to promote vivid narratives that encourage

Figure 6 The uncommon sight – according to recognition algorithms – of a cow on a beach.

us to learn from the mistakes and achievements of our shared past – should extend through its structures and into the human consciousness of the end user. If a technology frustrates such a fundamental goal, it is not fit for purpose.

Still, some knowledge will be held back from the public historical record. Some archives that deal with sensitive personal material discharge their duty by operating a dual public service whereby they make most collections public and offer private logins for other users to access material they have donated on the understanding that it should form part of the public record, but only once those it identifies are deceased. All the more crucial then, that the *boxes of stuff* be created in knowledge seekers' imaginaries, contributing a cultural confidence interval to their view of their subject and offsetting the clinical doubtlessness of the black box with a healthy dose of acknowledged ignorance.

Generally speaking, the struggle between archival thinking and computational thinking and the conceit of routine suggests a profound power shift in control of the knowledge we use to understand ourselves, question our assumptions and express our values in the most quotidian ways. Exclusion from meaningful engagement with processes shaping the technologies opening up their collections may result in a translation of cultural heritage specialists' complex knowledge of

material to a purely technical understanding of its existing functions. In order to prevent professionals in the social sciences and humanities from becoming completely alienated from powers that recontextualize not only their materials but their working practices, it has been argued that they must *socialize* the processes they have a stake in, rather than emulating the values of computational thinking.³⁷ Where fundamental questions of the power to name, classify and direct knowledge must be addressed, techno-solutionism's side-stepping should be resisted. Knowledge specialists' refusals to adopt black boxes whose input is a translation of their goals should be viewed in the context of other acts in defence of professional and disciplinary boundaries, which suggests high levels of receptiveness to the appropriate application of technology to solve discrete technical problems.

In summary, some sources of knowledge may be rendered effectively invisible by their exclusion from data-sharing initiatives, due to either practical constraints, wariness of the uncontrollable nature of the internet or a protective urge to shield it from the dark side of discoverability discussed earlier. Whatever the reason, the result is similar to preserving the solitary copy of a classic film on a DVD in a world with no remaining DVD players, or of a book in a language no living person can read. Knowledge flows are then diverted from reaching a wider audience, sometimes stymied by a lack of resources, which leads institutions to focus on inward-looking, short-term priorities rather than address long-term goals such as sharing, or the use of metadata that resists sharing due to incompatibility with standards used by other institutions or aggregators with whom institutions might otherwise share, or because those who are experts in a subject area and understand and support ways of sharing are inhibited by an institutional culture that does not reflect this open outlook. Where experts do have the power, they may have a particular reticence about the possibilities of data-linking. This may lead to the sharing of some data while other forms of sharing are delayed; this time lag alone is likely to shrink the profile of those sources, leading to a downward spiral in use.

Casting a shadow: A little sharing is a dangerous thing

In addition to the risk of knowledge invisibility through being left out of technologies for data sharing, another kind of invisibility operates through the use of these technologies. The construction of machine-learning algorithms involves the use of training data, classifications that those designing the

algorithm decide are appropriate to aid the algorithm's analysis, clustering of results and test queries that help engineers judge whether the algorithm is arriving at the 'right' answers. Human experts make decisions on which training data to use (usually those freely available in abundance, such as text from Amazon product reviews), how classifications are defined, whether and how to weight certain features, what test queries represent expected use, and what are the 'right' answers – all of which introduce bias and data voids into the process.³⁸ Algorithms may then be finalized when their results 'satisfice'.

The 'quick wins' of Google's immediacy and familiarity are a constant thorn in the side of practitioners concerned with upholding rigour in research methods, and there is a real fear that the celebration of openness is working as a diversion away from both the complex material excluded from it and any awareness that this phenomenon of hiddenness through eclipsing 'openness' is happening. It is clear that the new normal of the Google paradigm is having a direct effect on how knowledge seekers understand how to ask for knowledge, what timeframe and format of information is appropriate and desirable, and what constitutes a final result. Callon and Latour's description of a black box seems more pertinent than ever.³⁹ What is more, the coming together of the paradigms of the archival method and the computational method is viewed as imperilling archivists' fundamental values if the result is modelled on the algorithms of Google and Facebook, as described at a national library:

Even though people believe they see everything, they might see even less than before because they're only being shown the things that the algorithm believes they want to see. So, I'm really concerned with that increasing dominance of these organisations that commercial interests will increasingly drive knowledge creation ... [the reproductive functions of filter bubbles] make Facebook richer but society drifts apart much more and I would like to avoid that ... and not be stuck in something that looks a bit like Facebook large for cultural content.

(WP3 INT10)

It has also been suggested that widespread acceptance of machines doing a 'good enough' job would ultimately have a profound effect on the future of cultural heritage institutions by displacing the roles of human workers entirely from significant parts of cultural knowledge flows. Such displacement would represent a fundamental power shift, with the gatekeeping of cultural knowledge moving from specialists (concerned with content) to generalists (concerned with techniques). In this migration to the middle – training technologies on representative data, anticipation of likely queries and testing that presupposes

that the least wrong answer is the most right – it seems obvious that the most marginal knowledge will become further marginalized and minoritized as infrequent use is translated into minimal predictions of use. The unusual will become unused and ultimately unusable.

Big data has been conceptualized as a new medium⁴⁰ and, in that sense, we can see the potential for otherness to be marginalized in much the same way as television bombarded viewers with narratives of the nuclear family, reinforcing what was ‘normal’ while diminishing diversity through its relative obscurity of representation. The potential for big data, and particularly AI, to engender and reinforce disparity of visibility in any given setting is at least equal to conventional media.

An obvious example of (already) minoritized knowledge is knowledge about minorities. As we will discuss further in Chapter 7, minority languages are not well served by AI, which can lead to real-world minoritization as in Iceland, where the national language continues in spoken form but has been written out of instant text communication by indifferent tech firms for whom adapting their products makes no financial sense. Some knowledge sources present aspects of our cultural heritage that are under-represented because the collection and preservation of data relating to them were not prioritized as they were not valued or recognized in the context of a dominant narrative perceived to tell *almost* the whole story. These outliers are prized by researchers as their analysis provides the opportunity to make a unique contribution to knowledge (without which nobody would qualify for a PhD). The discovery of buried treasures, whose exploration through innovative methods and new perspectives enriches the historical record, demonstrates the value of archival practice. Such leaps are also the premise of the application of the techniques of big data.

Unfortunately, algorithms’ curiosity is more easily curtailed than that of human treasure hunters. As we saw in Chapter 4, search engine optimization (SEO) has shown how editing the content (data) and encoding (metadata) of a website exploits the easily pleased tendencies of algorithms. SEO incentivizes the creation of artificial links and the repetition of ‘keywords’ that are not necessarily key to the content in question, as well as spamdexing through wholesale reproduction of popular content. The power of attracting the attention of algorithms with specific triggers has been utilized by data activists. The principle of using *too much information* to flood data-processing systems has been weaponized in the cases of Hasan Elahi, who spared no detail of his mundane activities in offering up hourly updates to the US National Security Agency, as well as the initiative to ‘cc Theresa May’⁴¹ (then the UK Home

Secretary responsible for mass surveillance of electronic communications), among others. More regularly, hyper-visibility through SEO, although an industry in itself when taken to extremes, is considered essential to sharing knowledge in digital form.

Since academic journals became available online, authors have been asked to provide keywords that will make their work discoverable to knowledge seekers. Furthermore, the practice of quantifying the citations academic writing receives has been standardized and made searchable online as a tool for tracking implicit value – both of the scholar’s overall contribution to their field and of the individual text’s application to the work of others. Of course, researchers and more general users view, absorb and are informed by (i.e. *use*) many more materials than they directly quote or cite. If we were to view this fact as a *problem* in light of the potential for uncited material to be unfairly treated by a system designed to devalue data lacking in links in the form of references (compared to work that is controversial and therefore highly cited but does not make a positive contribution to scholarly discourse), we might come up with a *solution* similar to Amazon’s Kindle Popular Highlights. Kindle users were offered a feature that demonstrates the duality of Amazon’s business model as it uses its platforms to both advertise its substantive products and harvest the data of its users. The company explained that it was aggregating its customers’ highlighted passages to help readers focus on material that was meaningful to the greatest number of people.⁴²

Amazon’s interest in zeroing-in on what exact content appeals to its customers seems self-explanatory: people buy and read the same books for a multitude of reasons that do not always confirm a shared interest amongst them. In pinning down their interest to specific passages (containing keywords that can be cross-referenced across texts and easily recognized by algorithms), Amazon can classify readers in a much more sophisticated fashion, resulting in discrete, micro-targetable profiles. If such a model were to be extended to directly accessed digitized cultural knowledge, where might such a hunger for narrowing down *meaningfulness* take us? Finding out that the passages we highlight in a given text closely mirror the pattern of others’ highlights suggests to us that we are interested in that text for all the *normal* reasons, just as comparing one’s snaps of a city break to a composite image of the most photographed landmarks confirms that we chose a vantage point that is popularly agreed to confer recognizability. We might then ask what effect the overall *composition* of a knowledge resource has on the visibility of outliers.

Knowledge after Google: The agonism of archives and AI

Despite their role as *gatekeepers* of knowledge, cultural heritage practitioners are broadly supportive of what they see as the cosmopolitan, democratic spirit of cooperation embodied by data-sharing infrastructures. The fears they express are not based on power being put in the hands of knowledge seekers but that it ends up governed by corporate interests. When contemplating the paradigm shift of Technopoly – ‘the submission of all forms of cultural life to the sovereignty of technique and technology’ – Postman invokes Aldous Huxley’s categorization of Before Ford and After Ford eras.⁴³ Henry Ford is garlanded as the forefather of the production processes and capitalist values that have come to define societies in the industrialized world and his most famous apothegm is that his company’s Model T automobile was available in any colour ‘as long as it is black’. This restriction of choice, from a previous, wider range, represents an aspect of Ford’s production line efficiency (black paint dried more quickly than alternatives) that encapsulates concerns at the heart of both Huxley’s and Postman’s work.⁴⁴

Since George Ritzer introduced the concept of McDonaldization in 1993, many scholars have engaged with its applications to academic knowledge environments, most notably, libraries. Whereas some⁴⁵ have proposed alternative corporate strategies to counter the undesirable effects of following Ritzer’s model of *efficiency, calculability, predictability* and *control* (which lead libraries to a ‘simplistic’ approach that is ‘out of touch with the real needs of their users’ and fails to engage the ‘heads and hearts of their employees’⁴⁶), others have called for a wholesale rejection of *value* as determined by capital in favour of cultural heritage institutions’ *values* guiding their working practices.⁴⁷ To scholars of the social effects of new technologies, any discussion of Fordism or McDonaldization might evoke the scepticism Uricchio describes in reaction to apocalyptic headlines in which “new” technologies appear in the regalia of disruption.⁴⁸ Today’s datafied cultural heritage sector might be most usefully understood as contending with a *Starbuckization* – a phenomenon that takes account of the fact that even McDonald’s has shown flexibility in customizing its offering to a local audience, whereas Starbucks has exercised an even greater degree of control over its outlets (not franchises) and total standardization of its products. In displacing the local essence of a coffee shop with a globalized product, Starbucks has no truck with the comparatively adaptive model that affords the Spanish a beer with their Big Mac. Instead, Starbucks customers

the world over are choosing from exactly the same menu. This kind of cultural land-grab is analogous with the algorithmic regime, which, for Uricchio, is 'more than "just another" temporarily unruly new technology'.⁴⁹ What is notable about this analogy is that this claim to standardization is applied to seemingly inappropriate environments. Were these analogies not already too many cooks for the *data soup*⁵⁰ of cultural objects, descriptions and sources, we might consider the example of cookbooks released in the names of popular restaurants, a technology allowing the user to access the highlights of the institution at home, through the convenience of a self-guided experience that will inevitably fall short of the real thing as the chefs' expertise (as well as the full list of ingredients) is left out of the process. Much like the Starbucks offerings available onboard trains, with the promise of customer satisfaction comparable to the experience of having your coffee brewed to order by a barista, the added value of cultural heritage practitioners is undermined and erased.

The foregoing sections describe how big data disrupts knowledge creation and sharing where gatekeepers have some control over promoting the visibility of their data. Of course, there are many examples of developer priorities conflicting with the value of sharing, as in Google and Apple's exclusion of content from their platforms on the basis of commercial interests, judgements of inappropriate content or seemingly arbitrary decisions.⁵¹ The commercial logic of targeting customers by harvesting their data and offering them products likely to be of interest to them is clear. In providing users with a non-commercial service, there is little incentive to reject this prevailing logic: it is second nature for us to provide data to identify ourselves when using an interface now, whether by consciously adopting the data-double avatar that will represent us in our interactions on the platform or by passing through the checkpoint comprising two text boxes that expect our email address and password in order to passively browse a glorified shopping catalogue. Indeed, libraries have always required identification and their online checkpoints appear as an extension of this. Our browsing of sites like Amazon, however, is not exactly passive; through either logging in or accepting cookies, we are actively telling Amazon what we are interested in by letting it track our movements through the site. Without allowing ourselves to be counted in this way, ubiquitous features that now seem inseparable from the services of several web giants would not be possible. Other features have indubitable value beyond popular applications: while an Amazon-style linking to what *others* clicked on or a sort according to popularity over relevance may not suit the needs of a user searching for specific historical information, a Twitter-style sidebar trumpeting trending topics might provide a

useful signifier of relevance – appraising not the content that we have been given but the alignment of our interests to what others are seeking to engage with.

The websites, search engines and platforms provided by cultural heritage institutions do not collect collateral data like Google or Amazon do; they will therefore not be capable of performing predictive analytics, nor is it their task to develop AI ‘solutions’ to highly complex epistemological processes. The aim of cultural heritage institutions is not to intensify users’ engagement with their platform but to provide content appropriate to the research questions that are part of knowledge creation processes. Cultural heritage institutions therefore heavily rely on indexes, subject headings and controlled vocabulary and not on the search history of their users. By contrast, the big tech companies’ approach is marked by a ‘radical indifference’ towards content, which is merely ‘judged by its volume, range, and depth of surplus.’⁵² Moreover, knowledge seeking, which characterizes the activity of users of archives, libraries, museums and galleries, is not comparable to the consumption of cultural products, realized by Google through the selling of predictions of buying decisions and by Apple and Amazon by selling the product itself. Similarly, cultural heritage institutions are not interested in identifying patterns in knowledge-seeking procedures in the way that Amazon has harvested Kindle data.

Algorithms are generative⁵³ and they are used by the tech giants to stimulate ongoing use of their services. In this sense, Google has no interest in *solving* your query, that is, allowing you to find a simple answer to a simple question and put the matter to bed, any more than Facebook wants you to use its platform to passively check your closest friends’ statuses before logging out, satisfied that they are all doing fine. Without a click-through to some other step, AI does not learn: either to reinforce the supremacy of the most predictable association or to lend weight to a less common chain of content in a given context. The incentive, therefore, is to serve our interests in a way that makes more active participation irresistible. boyd notes how data presented to us in certain forms can make us more willing to override a quest for knowledge in favour of that which stimulates us, like a cheap high.⁵⁴ A feature of our addictive behaviour in response to such content is that we stay engaged, adding to chat about the viral meme to show that we got the joke or contributing to the debate on the controversial perspective to put right, out of a sense of justice, what we see as out of balance. Much attention is currently given to the issue of filter bubbles as spaces where conflicting views are not aired or engaged with but such bubbles would not grow, were they not pricked by an incendiary counter-view around which to rally. Furthermore, it has been suggested that being surrounded by

reinforcing opinions, whether intentional⁵⁵ or by dint of an algorithm learning your ‘preferences’, generates more radical views, behaviour or at least exposure to more radical content.⁵⁶

If we are to be thorough in our consideration of the characteristic principles of commercial data environments and what it might mean for the historical record if they were to be applied to cultural knowledge, we should also note the rationale of trending as a signifier of importance, the prioritization of literal currency but also the elevation of short-term relevance over continuing significance (on Twitter, for example, algorithms will minoritize the representation of issues that bubble under the surface for a prolonged period). Trends are by nature time-bound and consistency and continuity are devalued. It may indeed be a *myth* that those who work with machines and those who work with people represent separate cultures⁵⁷ but the agonistic struggle between computational thinking and archival thinking is undeniable.

Future invisibilities: Popular music, unmapped terrain and alternative facts

In the course of our research, people working in cultural institutions shared their fears for the future with us, resulting in (unprompted or perhaps *organic*) discussion of the ‘Googlization’ of the cultural knowledge currently safeguarded by conventional archival practices. What are we talking about when we talk about Googlization? The possibilities of data use that Google has realized through its insatiable ambition and its power to pour seemingly endless resources into projects that seize the value of previously unconnected knowledge fields have come to be taken for granted. Attempts to rival Google’s initiatives, such as Quaero (as a European alternative to Google Search) and French libraries’ attempts to digitize their own collections (as an alternative to Google Books), have been either abandoned completely or involved at least partial capitulation to Google due to relative lack of resources (Google has digitized the books of some of France’s public libraries in exchange for their rights). The Google model has nevertheless set a new bar, normalizing an omnipotent reach rather than specialization and a near instantaneous rate of output that has changed cultural expectations to such a degree that the ‘fast science’ methods of Surgisphere were not questioned until several peer-reviewed publications had already attained publication in some of the world’s most revered medical journals.

Googlization now seems inseparable from the customization of a service based on data about the user and to turn away from the advantages of such an approach is now as unthinkable as dismissing the utility of Henry Ford's production line. On a basic level, Googlization is concerned with the search engine itself and how the features we described in Chapter 4 become hegemonic. Google's algorithms also learn from its user base to deliver results that are likely to satisfy us, and results can be further personalized by favouring those that have suited similar users, based on what it knows about us, and those other users, from gathering data across its various platforms. The functionality allowed by such vast data banks normalizes the levels of personalization used to create a Spotify playlist or the businesses made visible to us on Google Maps. The duality of Googlized platforms affects what we will be offered when the machine is learning (songs representing two extremes of overlapping clusters of listeners so that we signal which musical path we would like to go down) and when it has figured out enough to satisfy us (leaving home improvement stores unmapped as it knows we live in rented property).

Aside from the risk of inaccuracies inherent in a blunt instrument pushing us to extremes to complete its learning objective or extracting paradata from us when we are not ready, applying a Googlized personalization simply runs counter to what drives most knowledge-seeking behaviour – the desire to be offered something we do not already know and were not likely to just stumble upon. Taking the example of Amazon's Kindle user data (which gobbles up half-formed ideas in Kindle notes even though we might erase them once we are better read), it is easy to see how common misunderstandings might gather pace in a post-truth era, creating a critical mass that could mean a new breed of fake news in the shape of machine-made conspiracy theories without even the intentionality of an original human source. On a more quotidian level, the creep of customization would favour confirmatory, rather than challenging, knowledge, putting ever more weight on the wording of a query. As AI is about prediction, we can expect to see something akin to a smartphone's autocomplete facility being incorporated into future experiences of navigating the historical record. Might we then treat the suggestions of an autocomplete function in the setting of a cultural heritage institution as metadata or even a result in itself? It would surely be fair to report on such a suggestion in the same way as the hackneyed reference to 'the *Oxford English Dictionary* definition ...', as this convention also reflects common usage. Data begets data. It is also worth noting that ubiquitous features of everyday technologies, such as autocomplete, are unlikely to register when applied to another context, but can seriously skew our

user experience. While in some settings that might lead to us drinking alone in the wrong Dog and Duck, or discovering we quite like two bands with very similar names, when searching for knowledge in an age of ‘alternative facts’, the stakes are higher.

Stjernfelt and Lauritzen found that entering ‘did the Hol’ into Google Search led to websites denying the Holocaust (they posit that serious studies do not tend to contain the same letter sequence, and so it would take deliberate SEO efforts to embed it in websites seeking to redress the balance by appearing on the first page of results alongside the conspiracy theorists).⁵⁸ Asking the right questions has always been important but in the era of fake news, starting with a shaky research question – or even the first few letters of one – could send you down a rabbit hole. Writing in 2019, danah boyd asks how far this rabbit hole might go, imagining the spiralling seduction of a slick perversion of ‘social justice’ taking a YouTube user from textbook definition to rightwing backlash in a matter of clicks.⁵⁹ A proponent of QAnon now sits in the US Congress and the spread of this outlandish conspiracy theory has demonstrated that the bar for the viability of claims gaining global traction is so low that any idea that simply putting verified facts out there would be a *fait accompli* looks incredibly naïve, and any attempt to limit access to knowledge in order to preserve elitist dissemination models starts to look dangerous.

boyd identifies the asking of questions such as ‘did the Holocaust really happen’ as the work of culturally induced ignorance studied by agnotologists. The production of ignorance is a political project, sometimes in the most literal way, as with colonial officials’ destruction of files that contained evidence of the British Empire’s atrocities, which they called Operation Legacy,⁶⁰ and sometimes more covertly. For example, colonial era records from the British occupation of Kenya, alongside which those that Operation Legacy reduced to ash would have formed a more complete historical record, now sit in the Foreign and Commonwealth Office’s archive at Hanslope Park. These ‘migrated archives’ are not publicly accessible (some are classified ‘top secret’, others are withheld from Freedom of Information requests with no reasons given), creating a data void as (un)touchable as the smoke that hung over New Delhi in 1957. The myth of the benevolent Empire continues to flourish in the classrooms and popular imagination of twenty-first-century Britain and boyd’s agnotological analysis of how newly created data voids can be exploited demonstrates just how seriously we should take both the purposeful production of ignorance and the practices that make it easier to reach falsehoods than knowledge that nourishes us. These are overt attacks on our culture.

Of course, the least we could hope for from a serious resource for seekers of cultural knowledge is that it would not peddle, or feed, conspiracy theories but it is worth considering the ramifications of transposing the taken-for-granted algorithms of one environment on another. Whether or not we consider ourselves to be under the cosh of Technopoly's push for efficiency, it seems unlikely that an algorithm would now be designed that deliberately did not seek to enhance users' experiences by correcting common mistakes. Machine prediction and enhancement now range across every medium of writing, from casual email,⁶¹ through Grammarly's business-speak services, to the short story suffering from writer's block,⁶² with Open AI's GPT-3 leading the way to new frontiers of sophistication, and bias. AI has now replaced humans in many media roles and, although academia is often described as an ivory tower, academic knowledge creation does not exist in a vacuum and is always, to some extent, informed by popular discourses – funding proposals are often framed by their relevance to current, headline-grabbing events for example, though they are always the product of years of prior research. Subtle shifts in the curation of news articles on aggregating platforms – whether we click through or scroll past – inform our perception of what matters to society at large.

Efficiency then manifests in the algorithmic equivalent of a library keeping its most used books by the front door – maybe even firing the text you have most likely come for into the revolving door for you to grab it without having to enter the building. Perhaps this would not be so bad, after all, when we had to rely on physical tomes, the inverse system ruled, leaving only the least popular knowledge available as that which most suited our needs had already been scooped up by fellow students with the wherewithal to plan their reading early. Rather than the reassuring certainty that we have been left with books no-one else is interested in, however, we cannot know for sure why the black-boxed algorithm has put these books by the door. We might be fairly confident that they are, in fact, the most used books in the library and, of course, this is a special library designed for us based on everything we have read before, but there is a Matthew effect in the black box that promotes the use of the most used. We might then worry about intellectual rigour if we are going to be diverted to *Existential Philosophy: A Reader*, rather than reading the works of Sartre directly but this is not the only level of accessibility on which the process discriminates. Complex knowledge forms will not have made it as far as the optimization process, meaning that only the data that is most accessible for machine learning will be selected for the shelves by the door – all the books that are deemed too deviant from the norm that has been designed for, simply

because they are too heavy, unusually shaped or containing unexpected media, are left in the basement stacks.

Hypernormalised hypermarkets of big data: Refusing to be cowed

Of course, if the result of the various kinds of invisibility discussed above was that we couldn't ever find what we were looking for, we would come to see such services as useless, stop using them and, if there really was no alternative, conclude either that the knowledge we seek cannot exist or that we must live under the oppression of ignorance as it is withheld from us.⁶³ What seems a more likely outcome, however – and is no stretch of the imagination for anyone with an ebay buyer's account – is that we will be the recipients of a product that *satisfices*. Pasquale describes Google and Apple as 'the Walmarts of the information economy, in that they 'habituate users to value the finding service itself over the sources of the things found'.⁶⁴ The invisibilization of provenance might be the most insidious effect of datafication because, when presented with irreconcilable knowledge claims, capacity to judge and choose between them will be diminished.

Following the model of the existing tech giants as a way to share knowledge (and at this point there is no possibility of a parallel knowledge market that signs up more of the global population than Facebook, re-maps our streets and closes its ears to Google searchers' questions based on the trends of Tweeters) cements invisibility not only through algorithmic practice. As long as there is commercialization, there will be an element of the secret sauce, not the transparency and accountability of a public service. Perhaps viewing the gatekeeping of knowledge as something not akin to ownership in line with capitalist principles was always naïve and we were enjoying the anomaly of a period of public ownership of knowledge only briefly. It is undeniably the case that academic journals have, until very recently, sought to keep knowledge from the general public by limiting access to their publications to those willing to pay a subscription themselves or, far more commonly, have the privilege of a university affiliation to draw on. The general public's grasp of the latest research was then dependent on what they were spoon-fed through reviews of academic literature in generalist publications and what served an often sensationalist media. Nevertheless, it has been argued that the 'secret DNA' of algorithms

should not be hidden from the public but open to public interrogation and discussion.⁶⁵ GDPR⁶⁶ has gone some way to providing for this.

We have seen the game-changing power of big tech: in calling time on presumptions of gender identity based on physical characteristics and in 2020, a year that will be remembered for racial reckoning in the United States, two of its 'Big Five' information technology companies (Amazon and Microsoft) stepped back from supplying facial recognition technology to police, and IBM withdrew from the technology, which has become synonymous with racial profiling, altogether. Pressure has been mounting on Facebook to take sustained action against campaigns of disinformation that led to the storming of the US Capitol in January 2021, precipitating its move to suspend Donald Trump's Facebook account. Pasquale's plea for Google to make its *secret DNA*, through which it *includes, excludes and ranks* – a power that dictates 'which public impressions become permanent and which remain fleeting'⁶⁷ – accessible to library scientists has so far gone unanswered.

The agonistic nature of the tension between computational thinking and archival thinking must be represented in the knowledge seeker's experience. The historical record is a process, not a product and so this dialogue between knowledge complexity and the technologies that might be applied to it should be ongoing. Presner describes this as not only a deconstruction of assumptions imposed on the cultural record but a move towards making space for 'the ambiguous, the unfinished, the differential, the multiple, and the spectral'.⁶⁸ We see this as not just living with the discomfort they engender but recognizing that discomfort as central to the process of the historical record and our most vital struggles to learn from it. For Star, being uncomfortable but 'content with that which is wild to us',⁶⁹ meant refusing the translation identified by Callon and Latour.⁷⁰ We argue that the obscuring practices of a *solutions* culture must be similarly refused in order to preserve untameable complex knowledge.

Leading the conversation with technical specialists and putting their perspectives on future knowledge seekers' needs in the foreground throughout should address practitioners' feelings of being on the back foot in responding to technologies as they become available to them or having to take up resources that do not fit existing needs because alternative models have not been worked up. The complexities of cultural data show us how shallow our knowledge environment might become if we lose sight of the rest of the 'knowledge iceberg' below the tip that big data approaches shine a spotlight on. The inherent invisibility introduced by data infrastructures' context-stripping standardization

processes demonstrates how sharing can come at a cost and what appear as freer flows of knowledge may be the result of a multitude of unseen dams. If our knowledge environment continues to develop according to the trajectories of datafication, without the check of critique that deconstructs the pervading assumptions of big data, what can we expect?

In a classic study,⁷¹ white children who had never seen a black child reading a book were shown a picture of a library and asked what the black child in the picture was doing (there was no black child in the picture). The white children's answers ranged between various menial tasks. Not one, students introduced to the study of racial stereotyping are told, said 'reading a book'. The white children's 'training data' had not equipped them to predict this minoritized material and the 'neutral' technology of a photograph only served to confirm their bias. We look to technology because we are aware of the (all too) human propensity to short-cut complexity: from misplaced keys to optical illusions to racial prejudice, we see what we expect (not necessarily want) to see. However, from misplaced cows to monocular illusions to racist predictions, we have seen that the black boxes of technology also short-cut complexity. We described in Chapter 5 how the context-stripping standardization processes of data infrastructures, which aggregate heterogeneous data, can add further layers of obfuscation, and we argue that the threat of obscurity this poses eclipses concerns about human error.

When considering the capacity of machine learning to contend with the art world, Wyse observes that neural networks' categorization certainty snowballs as images are tweaked to marginalize contextual 'noise'.⁷² Have we given enough thought to the consequences of using technology that takes Amazon product reviews as its starting point for knowledge? (The *intelligence* of such machine learning will be discussed in the next chapter). Of course, all forms of progress have to start somewhere, and advances such as those building on adversarial image research⁷³ help to expose vulnerabilities in the classification of knowledge that are perhaps fundamental to the drawing of such boundaries. Nevertheless, the problem of invisibility is not that the adoption of certain data practices, black-boxed systems, infrastructures and associated cultural norms temporarily covers over complex knowledge; it is that, in applying these layers, some of the noisy texture heretofore added by cultural practices is abraded so that holdings may fit together in a way that *makes sense*, and even if it is a more sophisticated technology that reaches back into the historical record to retrieve them, it is in this more governable, quieter form that they will be waiting.

Notes

- 1 Schwartz, Joan M. and Terry Cook, 'Archives, Records, and Power: The Making of Modern Memory', *Archival Science*, 2/1 (2002), 175.
- 2 Ibid., 176.
- 3 Joks, Solveig, Liv Østmo and John Law, 'The Infrastructures of Difference', *Nomadic Peoples*, 24/2 (2020), 323–43.
- 4 Latour, Bruno, *Science in Action: How to Follow Scientists and Engineers through Society* (Cambridge, MA: Harvard University Press, 1988).
- 5 Joks, Østmo and Law, 'The Infrastructures of Difference', 334.
- 6 Haraway, Donna, 'Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective', *Feminist Studies*, 14/3 (1988), 575–99.
- 7 Ibid., 590.
- 8 Star, Susan Leigh, 'Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions', *The Sociological Review*, 38/1 (1990), 26–56.
- 9 Van Schie, G., I. Westra and M.T. Schäfer, 'Get Your Hands Dirty in The Datafied Society', in *The Datafied Society: Studying Culture through Data* (Amsterdam: Amsterdam University Press, 2017), 187–8.
- 10 Pearl, J., 'The Limitations of Opaque Learning Machines', in *Possible Minds: 25 Ways of Looking at AI* (London: Penguin, 2019).
- 11 Bowker, G., 'Data Flakes: An Afterword to "Raw Data" Is an Oxymoron', in *'Raw Data' Is an Oxymoron* (Cambridge, MA: MIT Press, 2013), 168.
- 12 Nuffield Council on Bioethics, *Artificial Intelligence (AI) in Healthcare and Research* (London, 2018).
- 13 Barocas, Solon, Sophie Hood and Malte Ziewitz, *Governing Algorithms: A Provocation Piece* (Rochester, NY, 29 March 2013), <https://papers.ssrn.com/abstract=2245322>.
- 14 Kitchin, Rob, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London: SAGE).
- 15 Kitchin, Rob, 'Big Data, New Epistemologies and Paradigm Shifts', *Big Data & Society*, 1/1 (2014), 1–12.
- 16 boyd, danah and Kate Crawford, 'Critical Questions for Big Data', *Information, Communication & Society*, 15/5 (2012), 662–79.
- 17 Wang, H., 'Toward Mechanical Mathematics', in *The Modelling of Mind* (South Bend, Indiana: Bloomsbury, 1963), 93.
- 18 Postman, Neil, *Technopoly: The Surrender of Culture to Technology* (New York: Vintage, 1993), 70.
- 19 Economic and Social Research Council, 'Research Data Policy', *UK Research and Innovation* (2018), <https://esrc.ukri.org/funding/guidance-for-grant-holders/research-data-policy/>.

- 20 See, for example, <https://www.gwern.net>.
- 21 Piper, Kelsey, 'GPT-3, Explained: This New Language AI Is Uncanny, Funny – and a Big Deal', *Vox* (2020), <https://www.vox.com/future-perfect/21355768/gpt-3-ai-openai-turing-test-language>.
- 22 Kitchin, Rob, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences* (London, 2014), 22.
- 23 Williamson, Ben, 'Political Computational Thinking: Policy Networks, Digital Governance and "Learning to Code"', *Critical Policy Studies*, 10/1 (2016), 39–58.
- 24 Merton, Robert K., 'The Matthew Effect in Science | Science', *Science*, 159/3810 (1968), 56–63.
- 25 Edmond, J., M. Priddy and N. Horsley, 'KPLEX Report on Historical Data as Sources' (2018), <https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:114127>.
- 26 O'Neil, C., *On Being a Data Skeptic* (Sebastopol: O'Reilly Media, 2014), 2.
- 27 Latour, Bruno, *Science in Action: How to Follow Scientists and Engineers through Society* (Cambridge, MA, 1988).
- 28 Layne, Linda L., 'The Cultural Fix: An Anthropological Contribution to Science and Technology Studies', *Science, Technology, & Human Values*, 25/3 (2000), 352–79.
- 29 Zax, David, 'Siri, Why Can't You Understand Me?', *Fast Company* (2011), <https://www.fastcompany.com/1799374/siri-why-cant-you-understand-me>.
- 30 'When It Comes to Gorillas, Google Photos Remains Blind | WIRED', <https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/>.
- 31 Lyons, Kim, 'Google AI Tool Will No Longer Use Gendered Labels Like "Woman" or "Man" in Photos of People', *The Verge* (2020), <https://www.theverge.com/2020/2/20/21145356/google-ai-images-gender-bias-labels-people>.
- 32 American Civil Liberties Union (2018), <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28>.
- 33 Snow, Jacob, 'Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots', *American Civil Liberties Union* (2018), <https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28>.
- 34 Buolamwini, Joy and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', in *Conference on Fairness, Accountability and Transparency* (presented at the Conference on Fairness, Accountability and Transparency, 2018), 77–91, <http://proceedings.mlr.press/v81/buolamwini18a.html>.
- 35 Beery, Sara, Grant Van Horn and Pietro Perona, 'Recognition in Terra Incognita', in *Computer Vision – ECCV 2018*, ed. Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu and Yair Weiss, Lecture Notes in Computer Science (Cham, 2018), 472–89.

- 36 Lyons, 'Google AI Tool Will No Longer Use Gendered Labels like "Woman" or "Man" in Photos of People'.
- 37 Ruppert, Evelyn, 'Rethinking Empirical Social Sciences', *Dialogues in Human Geography*, 3/3 (2013), 268–73.
- 38 boyd, d., *It's Complicated* (New Haven and London, 2014); Caliskan, Aylin, Joanna J. Bryson and Arvind Narayanan, 'Semantics Derived Automatically from Language Corpora Contain Human-like Biases', *Science*, 356/6334 (2017), 183–6.
- 39 Latour, *Science in Action: How to Follow Scientists and Engineers through Society*.
- 40 Lushetich, N., *Big Data – A New Medium?* (Abington: Routledge, 2020).
- 41 As Home Secretary, Theresa May championed legislation popularly dubbed the Snooper's Charter, requiring internet and mobile service providers to keep records of internet usage, voice calls, messages and email for up to a year in case police requested access to the records while investigating a crime. Civil servant Eric Finch launched the campaign 'National cc all your e-mails to Theresa May Day' on 1st May 2012, as a public Facebook page, in protest.
- 42 Striphas, T., 'How to Have Culture in an Algorithmic Age', *The Late Age of Print*, (2010), <https://www.thelateageofprint.org/2010/06/14/how-to-have-culture-in-an-algorithmic-age/>.
- 43 Postman, Neil, *Technopoly: The Surrender of Culture to Technology* (New York, 1993), 49–52.
- 44 Huxley, Aldous, *Brave New World* (London: Vintage, 1932); Postman, *Technopoly*.
- 45 Quinn, B., 'The McDonaldization of Academic Libraries?', *College and Research Libraries*, 61/3 (2000), 248–61.
- 46 Ibid., 254.
- 47 Fister, B., 'Librarians as Agents of Change, Keynote Address', in *Libraries Out Loud: New Narratives of Enduring Values* (presented at the ACRL Oregon-Washington Joint Fall Conference, Corbett, Oregon, 2012); Nicholson, Karen P., 'The McDonaldization of Academic Libraries and the Values of Transformational Change | Nicholson | College & Research Libraries', <https://crl.acrl.org/index.php/crl/article/view/16428>.
- 48 Uricchio, W., 'Data, Culture and the Ambivalence of Algorithms', in *The Datafied Society: Studying Culture through Data* (Amsterdam: Amsterdam University Press, 2017), 125.
- 49 Ibid.
- 50 Edmond, Jennifer, Natasa Bulatovic and Alexander O'Connor, 'The Taste of "Data Soup" and the Creation of a Pipeline for Transnational Historical Research', *Journal of the Japanese Association for Digital Humanities*, 1 (2015), 107–22.
- 51 Pasquale, Frank A., *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge: Harvard University Press, 2015).

- 52 Shoshana Zuboff, *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power* (New York: PublicAffairs, 2019), 505.
- 53 Lash, Scott, 'Power after Hegemony: Cultural Studies in Mutation?', *Theory, Culture & Society*, 24/3 (2007), 55–78.
- 54 boyd, danah, 'Streams of Content, Limited Attention: The Flow of Information through Social Media' (2009), <https://er.educause.edu/articles/2010/10/streams-of-content-limited-attention-the-flow-of-information-through-social-media>.
- 55 Sunstein, C. R., *Going to Extremes: How Like Minds Unite and Divide* (Oxford: Oxford University Press, 2009).
- 56 Tufekci, Zeynep, 'YouTube, the Great Radicalizer', *The New York Times* (10 March 2018), section Opinion, <https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html>.
- 57 Star, 'Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions'.
- 58 Stjernfelt, Frederik, and Anne Mette Lauritzen, *Your Post Has Been Removed: Tech Giants and Freedom of Speech* (2020), <https://www.springer.com/gp/book/9783030259679>.
- 59 boyd, danah, 'Agnotology and Epistemological Fragmentation', *Medium* (2019), <https://points.datasociety.net/agnotology-and-epistemological-fragmentation-56aa3c509c6b>.
- 60 Cobain, Ian, *The History Thieves: Secrets, Lies and the Shaping of a Modern Nation* (London: Portobello Books, 2016).
- 61 Seabrook, John, 'Can a Machine Learn to Write for The New Yorker?', *The New Yorker* (2019), <https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker>.
- 62 Epstein, Sophia, 'How Do You Control an AI as Powerful as OpenAI's GPT-3?', *Wired UK* (27 July 2020), <https://www.wired.co.uk/article/gpt-3-openai-examples>.
- 63 Hypernormalisation (2016), [documentary] Dir. Adam Curtis, UK: BBC.
- 64 Pasquale, *The Black Box Society*, 85.
- 65 Stjernfelt, Frederik and Anne Mette Lauritzen, *Your Post Has Been Removed: Tech Giants and Freedom of Speech* (2020), <https://www.springer.com/gp/book/9783030259679>, 164.
- 66 'General Data Protection Regulation (GDPR) – Official Legal Text', <https://gdpr-info.eu/>
- 67 Pasquale, *The Black Box Society*, 61.
- 68 Presner, Todd, 'The Ethics of the Algorithm: Close and Distant Listening to the Shoah Foundation Visual History Archive', in *Probing the Ethics of Holocaust Culture* (Cambridge, 2016), 175–202.
- 69 Star, 'Power, Technology and the Phenomenology of Conventions: On Being Allergic to Onions', 82.

- 70 Latour, *Science in Action: How to Follow Scientists and Engineers through Society*.
- 71 Note: We can find no reference for this study; perhaps it is a sociological myth – or perhaps our keyword search was ineffectual because Google flagged the content we were looking for as too controversial to return a result.
- 72 Wyse, L., ‘Appreciating Machine-Generated Artwork through Deep Learning Mechanisms’, in *Big Data – A New Medium?* (Abingdon: Routledge, 2020).
- 73 For example, Su, Jiawei, Danilo Vasconcellos Vargas and Sakurai Kouichi, ‘One Pixel Attack for Fooling Deep Neural Networks’, *IEEE Transactions on Evolutionary Computation*, 23/5 (2019), 828–41.